PERCLOS (CVI-TRUCK)

U.S. Department
of Transportation
National Highway
Traffic Safety
Administration

NHTSA
People Saving People
htcp://www.nhtsadot.~ov

DOT HS 808 762

April 1998

Final Report

Evaluation of Techniques for Ocular Measurement as an Index of Fatigue and the Basis for Alertness Management

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT HS 808 762		

4. Title and Subtitle	5. Report Date
Final Report: Evaluation of Techniques for Ocular Measurement as an Index of Fatigue and as the Basis for Alertness Management	April 1998
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
David F. Dinges, Ph.D., Malissa M. Mallis, Greg Maislin, MA, MS., John Walker Powell, IV, M.A.	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
National Highway Traffic Safety Administration	11. Contract or Grant No. DTNH22-93-D-07007

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
National Highway Traffic Safety Administration 400 Seventh Street, S.W. Washington, DC 20590	NHTSA Contractor Report
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This final report establishes the scientific validity of the ocular measure "Perclose" as a generally useful and reliable index of lapses in visual attention, i.e. the percentage of eyelid closure over the pupil. Perclose was previously specified as a relevant measure of drowsiness in several driving simulator studies (NHTSA Final Report, DOT HS 808 640). In the present research further validation of Perclose was established among other measures, in a controlled sleep deprivation study, using a well-known psychophysical index of lapses in visual attention, i.e., Psychomotor Vigilance Task (PVT). The present study provides the scientific and practical basis to relate real-time lapses in visual attention to over-the-road driving performance.

17. Key Words	18. Distribution Statement
Drowsy Driving, Fatigue, Driver Monitoring, Driving Simulation, Vigilance, Driver Impairment, Drowsiness Detection, Technology Validation, Psychophysical Measurement	Document is available to the public through the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20.Security Classif. (of this page)	21. No of Pages	22. Price
Unclassified	Unclassified	TBD	

Form DOT F 1700.7(8-72) **Reproduction of completed page authorized**

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

I. EXPERIMENT ON PERFORMANCE-BASED VALIDATION OF TECHNOLOGIES

in recent years, an increasing number of drowsy driving biobehavioral monitors have become available. There is a widespread hope that such technologies will be a key component in effective management and prevention of drowsy driving. To obtain estimates of the current scientific validity of six promising drowsiness-detection technologies, a double-blind, controlled laboratory validation experiment was undertaken. The six technologies tested included a video-based scoring of eye closure by trained observers; two EEG algorithms; a head tracker device; and two wearable eye-blink monitors. The six technologies yielded a total of nine drowsiness metrics. Psychomotor vigilance task (PVT) performance lapses were selected as the validation criterion variable for the technologies for three reasons: (1) driving is fundamentally a vigilance task requiring psychomotor reactions; (2) psychomotor vigilance has been validated to be very sensitive to fatigue from night work and sleep loss; (3) hypovigilance while driving is the outcome most fatigue-detection technologies seek to identify.

Fourteen healthy adult males remained awake in the laboratory for 42 hr, while working on a computerized test battery every 2 hr that included a 20-min. PVT task. PVT performance lapses were recorded each minute throughout each PVT trial, and totaled for the entire 20 minutes. Each technology was time-locked to PVT performance to test coherence between vigilance lapses and each technology's specific drowsiness metric. Thus, in order for a technology to demonstrate high coherence with PVT lapses, it's drowsiness metric had to demonstrate systematic covariation with performance lapses across the 42-hr period of waking. Drowsiness metrics were obtained from each technology source, but the sources remained blind

both to the lapse data and to the specific hour of continuous wakefulness at which each data were acquired, to ensure unbiased estimates of impairment.

Although complete data for any given technology were often only available on a subset of the 14 subjects, the results of the experiment were relatively straight forward. Nearly ail of the technologies showed potential for detection of drowsiness-induced hypovigiiance by accurately predicting lapses in at least one subject or a subset of subjects. Only one technology, however, PERCLOS-the video-based scoring of slow eye closure by trained observers (Wierwiiie et al., 1994; Wierwille & Ellsworth, 1994)--correlated highly with PVT lapses both within and between subjects (mean $r = 0.875$, $p = 0.00001$ for lapse frequency; mean $r = 0.919$, $p = 0.00001$ for lapse duration [See Appendix]). Meeting the validation criterion both through high intra-subject and high inter-subject coherence is an important and highly promising outcome, since one of the more serious problems plaguing fatigue-detection and prevention is the large inter-subject differences in vulnerability to fatigue, such as was seen in the recent USA-Canada driver fatigue and alertness study (Wylie et al., 1996). PERCLOS not only had the highest coherence of the technologies tested, but it also correlated more highly with PVT lapses than did subjects' own ratings of their sleepiness ($t = - 3.9$, $p = 0.003$), and it had high positive and negative predictive values. It was also clear that PERCLOS was predictive of lapses in the first 22-hr of waking (mean $r = 0.872$)--a time frame that should apply to the majority of drivers.

The three drowsiness metrics of PERCLOS consistently covaried with PVT lapses across the 42-hr, but there was no difference among them in accuracy. However, the time base used to calculate drowsiness for PERCLOS and all the other technologies markedly influenced the coherence with PVT lapses. Essentially all technologies performed better when predicting lapses

over a 20-minute period than over a 1 -minute period. PERCLOS 1 -minute coherence scores were well above those of other technologies (mean $r = 0.768$, $p = 0.0001$ for lapse duration).

To the extent that each technology was capable of predicting PVT lapses from at least one subject, all of the technologies tested displayed some degree of potential as hypovigilance detectors. Therefore it remains possible for some of these technologies to further improve their "detection" of lapses. With this in mind, PVT lapse data have been sent to each supplier (after the results of this prospective study were established), to permit a retrospective "tweaking" of drowsiness algorithms for the purpose of enhancing their detection of lapses. However, a cautionary note is in order. If such a retrospective "enhancement" of coherence proves possible for some of the technologies, a prospective re-validation test of the "tweaked" drowsiness algorithm would be necessary.

The results of this experiment in conjunction with the work of Wierwille et al. (1994) suggest that PERCLOS has the potential to detect fatigue-induced lapses of attention during driving, if the following can be achieved. (1) The PERCLOS scoring algorithm used by human observers in laboratory studies must be automated in a computer algorithm with demonstrated evidence of acceptable levels of validity and reliability. (2) PERCLOS must be validly and reliably measured during driving, using unobtrusive technologies (e.g., video image analysis, infrared eye tracking). (3) Acceptable levels of positive and negative predictive values for driver fatigue must be determined for an automated, over-the-road version of PERCLOS. Attempts to meet the above criteria in order to transition an on-line, automated version of PERCLOS to a **realistic** over-the-road environment are currently underway.

II. EXPERIMENTAL STUDY OF EFFECTS OF ALERTING STIMULI

A second experiment was a pilot study derived from the first experiment and concerned the effects of alerting stimulation on drowsiness-induced PVT lapses. The deployment of on-line driver monitoring technologies to prevent drowsy driving may involve a drowsiness-detection system coupled to an alerting stimuli, to not only warn the driver of hypovigilance but to also help the driver overcome the drowsiness long enough to depart the roadway and rest.

In a second experiment derived from the first study, four subjects repeated the validation experiment but this second time, throughout the 42-m period of waking, they were exposed to both auditory and vibrotactile alerting stimuli applied during the PVT performance trials.

Auditory and vibrotactile alerting stimuli were delivered during each 20-minute PVT performance bout in Experiment II, based on the average lapse profile of subjects in Experiment I (e.g., more stimuli were delivered at times when lapsing was elevated in Experiment I, such as in the middle of the night). Thus, although alerting stimuli were not contingent on actual PVT lapses (there were a number of reasons why this was neither practical nor theoretically optimal), alerting stimuli did increase in frequency and diversity as lapsing would be expected to characteristically increase across hours awake and time of day. Each vibrotactile stimulus was delivered through the hand-held PVT response box. Auditory stimuli consisted of three different pre-recorded messages presented by a female voice (“stay awake, stay alert,” “watch for the stimulus,” and ‘please pay attention”).

Comparisons of PVT lapses in each minute prior to, during, and following individual and collective stimulation revealed that providing auditory + vibrotactile alerting stimuli did not markedly reduce lapses in drowsy subjects beyond the minute in which the alert occurred. A parallel analysis of PERCLOS confirmed this finding. These results suggest that a study of the

effects of alerting stimuli on drowsy drivers should focus on the most robust combination of alarm and the most potent alerting stimuli. There is some evidence that certain olfactory, thermoregulatory, and social stimuli may possess more potent alerting potential than auditory and vibrotactile stimuli. These stimulus modalities should be the focus of future research. Regardless of the duration of the acute effects of alerting stimuli, there is also a need to determine how drowsiness alarms and alerting stimuli are used by drivers under time pressure.

ACKNOWLEDGMENTS

The research and substantive evaluation upon which this article was based were supported primarily by contract DTNH22-93-D-07007 from the U.S. Department of Transportation, which included support from the Federal Highway Administration--Office of Motor Carriers (I. Experiment on Performance-Based Validation of Technologies), and support from the National Highway Traffic Safety Administration (II. Experimental Study of Effects of Alerting Stimuli). Additional support for the projects was provided in part by grant F49620-95-l-0388 from the U.S. Air Force Office of Scientific Research, by cooperative agreement NCC-2-599 from U.S. National Aeronautics and Space Administration, by grant NR04281 from the National Institutes of Health, U.S. Public Health Service, and by the Institute for Experimental Psychiatry. We thank the volunteer subjects for contributing their time and best efforts in the conduct of these experiments. We appreciate the excellent cooperation provided at many levels by the suppliers of the drowsiness-detection technologies. We acknowledge the following individuals for their assistance in performing this research: Pawel Adrjan, Barbara R. Barras, Julie Buxbaum, Michele M. Carlin, Natalie Denney, . Christine M. Dinges, Kelly A. Gillen, Dr. Richard Grace, Robert Hachadoorian, Cathy Hwang, Beatrice Jauregui, Cristian Jurau, Nicole Konowol, Jennifer Law, Donald Luong, Ravi Mariathason, Matthew Martino, Jennifer McKenna, Raj Mittal, Lan Nguyen, Emily Carota Ome, Jason Parkin, Sheelu Samuel, Dr. James Staszewski, David Thakker.

I. EXPERIMENT ON PERFORMANCE-BASED VALIDATION OF TECHNOLOGIES

SPONSORED BY: FEDERAL HIGHWAY ADMINISTRATION--OFFICE OF MOTOR CARFUER~

INTRODUCTION

Although scientific and applied initiatives to develop on-line measures of alertness/drowsiness and hypovigilance have a long history (O'Hanlon & Kelley, 1974; Dinges & Graeber, 1989; Brookhuis, 1995), this area has undergone renewed interest and intensified activity especially in the USA (Rau, 1996; Knipling, 1996), and in Europe (Brown, 1995, 1997) in the past 3 years (Dinges, 1995a,b, 1996,1997). There are four primary reasons for this "investment" in technology to manage fatigue-related performance impairment in all modes of transportation, but especially in commercial motor vehicle operations (Dinges & Mallis, in press).

(1) Fatigue-related crashes are common and serious. Since 1994, there has been growing evidence from industrialized countries that fatigue from varying combinations of sleep loss, night driving (i.e. circadian rhythms), and prolonged work time (i.e., wake time on task) contributes to substantial numbers of motor vehicle crashes. Although estimates and the methods on which they are based vary widely, few dispute that the problem of drowsy driving is inadequately addressed. Fatigue has been estimated to be involved in 2% to 23% of all crashes (cf., O'Hanlon, 1978; McDonald, 1984; Home & Reyner, 1995; Knipling & Wang, 1995; Maycock, 1997); in 4% to 25% of single-vehicle crashes (cf., Wang & Knipling, 1994; Brown, 1995); in 10% to 40% of crashes on long motor ways (Shafer, 1993; Dinges 1995b); and in 15% of single-vehicle fatal truck crashes (Wang & Knipling, 1994). Fall asleep crashes are also very serious in terms of

injury severity (Pack et al., 1995). In the USA, fatigue has been implicated as the most frequent contributor to crashes in which a truck driver was fatally injured (U.S. National Transportation Safety Board, 1990). Much of the focus on the putative role of sleepiness/drowsiness in traffic accidents has centered on single vehicle crashes, although there is no reason to believe that sleepiness is not also involved in multiple vehicle crashes. In the USA, single vehicle crashes in which no alcohol was involved account for more than a quarter of all fatal crashes, more than a quarter of all injury-only crashes, and more than a quarter of all property damage-only crashes (see Dinges, 1995b). While many factors can contribute to single vehicle non-alcohol-related crashes, the fact that they comprise 27% (i.e., 1.67 million crashes in 1993) of all motor vehicle crashes in the USA suggests that fatigue may contribute to more of these crashes than current estimates allow. Consequently, with so many persons driving fatigued, technologies that detect dangerous levels of sleepiness before a crash occurs are essential.

(2) Subjective estimates of sleepiness are unreliable. Experiments have demonstrated that subjects cannot reliably predict when they are impaired to the point of having an uncontrolled sleep attack (i.e., microsleep) and/or a serious vigilance lapse (Dinges, 1989). Drivers know when they are experiencing sleepiness (Home & Reyner, 1995), but they cannot necessarily translate those introspections into accurate predictions of how long their eyes are closed and whether they are missing signals, or when they will have an uncontrolled sleep onset while driving (Wylie et al., 1996; Brown, 1997). On the other hand, although self-reports of sleepiness are highly influenced by contextual variables (Dinges, 1989,1995b), drivers should know when they are experiencing heavy eyelids and head bobbing, which is likely past the point of impairment by drowsiness (Kribbs & Dinges, 1994). Hence, technology may offer the potential

for an earlier and more reliable warning of performance-impairing sleepiness, before drowsiness leads to a catastrophic outcome.

(3) Drowsiness-detection technology may offer an alternative to proscriptive hours of service. Technology is viewed by some as a key component in a package of fatigue management options that can replace or at least put flexibility into federally-mandated proscriptive hours of service. For example, the current USA federal hours of service for commercial motor vehicle operators were written in 1939, and rely on work time as the primary determinant of fatigue (this is not unique to the trucking industry). It has been recognized for some time, however, that within limits, work duration accounts for only a modest proportion of accident risk (Hamelin, 1987). Thus the current hours of service may not prevent many fatigue-related crashes. even when compliance is 100%. Fall-asleep crashes are more likely to occur during night driving and in sleep-deprived persons (e.g., Harris, 1977; Mitler et al., 1988; Pack et al., 1995). This is consistent with scientific studies in the past 30 years that have demonstrated that the level of waking alertness is regulated by two neurobiological forces that shape the time course of subjective fatigue and aspects of performance--the endogenous circadian rhythm and the need for sleep (Dinges, 1995b). When considered together and in combination with work hours, the product of these processes regulating fatigue and vigilance is nonlinear, temporally dynamic, and complex. This makes it complicated to derive regulatory schemes to prevent fatigue. Hence, technologies that monitor the driver's temporally dynamic state of alertness/drowsiness over time are viewed as offering an advantage over proscriptive regulations. Such technologies may provide one of a number of ways to optimize safety through preyention of fatigue-related crashes, while permitting greater flexibility in work-rest scheduling to facilitate economic and related pragmatic goals, as well as drivers' personal choices.

(4) Technological advances have made the goal feasible. Many technologies being developed for detecting drowsiness are miniaturized and unobtrusive (their durability and cost-effectiveness are less well established). Advances in electronics, optics, sensory arrays, data acquisition systems, algorithm development (e.g., neural nets), and other areas have made it far more likely that the goal of an affordable drowsiness-detection system in a truck or automobile will be achieved and implemented in far less than the 10-20 years estimated by Brown (1995, 1997). In the USA, for example, there are currently many efforts underway at federal, industry, and entrepreneurial levels toward development of technologies for monitoring a driver's physiology or behavior in order to "manage" performance-impairment from fatigue in transportation (Dinges & Mallis, in press). This marriage of technology and the human operator for drowsiness detection is part of a broader emphasis in the USA on development of "intelligent vehicle" and "driver condition warning" initiatives. Thus, technology development for managing driver fatigue extends beyond on-line devices that monitor the operator during the driving task, and broadly includes "readiness-to-perform and fitness-for-duty technologies," "mathematical models of alertness dynamics joined with ambulatory technologies;" and "vehicle-based performance technologies" (Dinges, 1997; Dinges & Mallis, in press).

OPERATOR-CENTERED, IN-VEHICLE, FATIGUE-MONITORING TECHNOLOGIES

Operator-centered, in-vehicle, fatigue-monitoring technologies seek to record some biobehavioral dimension(s) of an operator, such as a feature of the eyes, face, head, heart, brain electrical activity, etc., on-line during driving (Dinges, 1995a, 1997; Dinges & Mallis, in press). An ongoing review and categorization of technologies in this area reveals that there are currently more than 20 different initiatives in on-line biobehavioral monitoring in various stages of development (Mallis & Dinges, in preparation). All of the biobehavioral monitoring technologies

currently being proposed to continuously record driver alertness, drowsiness or vigilance capability are in the prototypical development, validation testing, or early implementation stages. Their full effectiveness, implementation, and acceptance remain unproven scientifically and practically (Dinges & Mallis, in press), which is problematic given the growing support for technology development in fatigue management (the facilitators), the entrepreneurial zeal currently overtaking technology companies in this area (the suppliers), and the escalating attractiveness of fatigue management technologies to transportation industries (the buyers). There is a risk of a rush toward widespread use of technologies that do not validly or reliably detect fatigue. If fatigue-monitoring technology development continues and is proposed as one piece of a programmatic "fatigue management" alternative to proscriptive hours-of-service regulations, then technologies that are alleged to be effective must be shown to meet or exceed a range of criteria involving scientific, practical, and legal/ethical standards (Dinges, 1995b, 1996, 1997). A great deal of harm can be done if invalid and/or unreliable devices are quickly and uncritically implemented. In addition to the potential for increased crash risk, deployment of invalid and/or unreliable fatigue-detection technologies will result in wasted resources and provide only a false sense of security and fatigue management. Scientific validity is the first standard a technology must meet (Dinges, 1995a,b, 1996, 1997; Dinges & Mallis, in press).

SCIENTIFIC VALIDITY OF DROWSINESS-DETECTION TECHNOLOGIES

This study was concerned with the scientific validity and reliability of a number of the more promising operator-centered, fatigue-detection, technologies. Most technologies explicitly claim or imply detection of some aspect of either a heightened risk of operator error or outright impairment through one or more of the following hypothetical constructs: operator vigilance; operator attention/inattention; operator alertness/drowsiness; operator microsleeps; operator

hypovigilance; operator performance variability; or operator vulnerability to error. Devices that purport to detect a fatigued operator, therefore, must demonstrate that they detect some aspect of fatigue/drowsiness/hypogilance relevant to driving performance (the validity standard), and that this detection is repeatable (the reliability standard). Even if a device is valid and reliable, to be practically useful, it must meet additional standards of high sensitivity and high specificity. Thus, the device must detect all (or nearly all) fatigue events and fatigued operators (i.e., high sensitivity standard), without too many false alarms (i.e., high specificity standard). A device that has high sensitivity but low specificity may detect hypovigilance, but may give too many false alarms to be useful. In contrast, a device with low sensitivity but high specificity may give few false alarms, but it may miss too many hypovigilance events to be useful (Dinges, 1997; Dinges & Mallis, in press).

METHODS

STUDY DESIGN

The design used permitted an estimate of the coherence between psychomotor vigilance test (PVT) performance lapses (i.e., the validation criterion variable) and the alertness/drowsiness output of each technology/algorithm within a given individual subject. This was possible by varying the endogenous level of alertness/drowsiness across a broad range, within each subject, through 42-hr of objectively-documented, sustained wakefulness (sleep deprivation) in the laboratory, and by recording PVT performance and concomitant technology/algorithm outputs every 2 hr throughout this 42-hr period. The degree of coherence manifesting in specific technologies/algorithms was conceptualized as the degree to which each could reproduce the relative alertness/drowsiness orderings of discrete time periods as determined by the validation criterion (i.e., PVT lapses) calculated within each subject, across the 42-hr period. (The reason

for using the term “coherence” to describe the relationship between each technology’s drowsiness index and psychomotor vigilance lapses is explained later in the “Statistical Approach” section). The resulting coherence metrics were summarized to determine “average coherence” and then compared among subjects to determine the extent to which each given technology/algorithm yielded consistent degrees of coherence across different individuals. Four control procedures were added to the design to enhance the integrity of the findings.

(1) Each technology/algorithm was time-locked in real time to PVT performance to permit coherence estimates for minute-to-minute fluctuations and bout-to-bout fluctuations in alertness-drowsiness across the 42-hr period of wakefulness.

(2) The suppliers of technologies were blind to PVT lapse data (i.e., the criterion variable) during the course of their extracting drowsiness/alertness scores from their technology/algorithm, while investigators at the University of Pennsylvania were blind to each technology’s scoring algorithm. This double-blind procedure was maintained throughout data acquisition and analyses. Without knowledge of subjects’ PVT lapses, suppliers of technologies were required to furnish at least one alertness/drowsiness score for each minute and for the entire 20-min. PVT bout for each subject on which their technology/algorithm was applied. Thus, the study design permitted a double-blind, prospective test of the coherence between PVT lapses and each technology/algorithm. This approach was deemed optimal for preventing unwitting bias from contaminating coherence estimates.

(3) To optimize the reliability of coherence estimates further, technology suppliers were also kept blind to the timing of data acquisition. Investigators at the University of Pennsylvania provided suppliers with their own specific technology/algorithm raw data organized in files that were randomly coded for temporal order (except for bout 1, which was allowed to permit

suppliers to “calibrate” their algorithm for each subject when the subject was fully alert). In other words, each technology supplier knew which 20 data files were associated with a given subject, and which file was the first test trial for that subject (i.e., bout 1 at 10:00 a.m. on day l), but each supplier was blind to the timing of the remaining 19 data bouts for that subject. This procedure prevented bias based on knowledge of the length of time subjects had been awake (and/or time of day) from influencing supplier data processing and extraction of the 1 -minute and 20-minute (bout) alertness-drowsiness metric(s).

(4) Processed data (drowsiness scores) received from technology suppliers, and PVT lapse data (criterion vigilance performance scores) from the University of Pennsylvania were electronically forwarded to an independent professional statistician for calculation of coherence results. Suppliers of technology were kept blind to the results from all other suppliers.

SUBJECTS

Fourteen healthy adult male subjects (ages 21-39 years) were studied in the laboratory during 42 hr of total sleep deprivation (TSD). Volunteers were asked to participate based on their expressed interest, availability, and health. Female subjects were solicited but none volunteered, which is consistent with our experience performing sleep deprivation experiments in the past 20 years--many more males volunteer for such studies than females. All subjects were healthy, nonsmokers who consumed no more than an average amount (i.e., 500 mg or less) of caffeine per day. Subjects were also screened to ensure they had stable sleep/wake cycles and that they were free of sleep disorders. Table 1 displays the subjects’ characteristics and sleep tunes.

All aspects of the experimental protocol, procedures, and informed consent were approved prior to initiating any investigation with subjects, by the University of Pennsylvania

Committee on Studies Involving Human Subjects. The 42-hr TSD protocol was briefly described in public advertisements posted in local newspapers and on campus bulletin boards to recruit

Table 1. Characteristics of subjects studied.

ID	Age (y)	Gender	Height	Weight (lb.)	Mean total sleep time (TST) 4 days prior to study (hr)	TST (hr) night before study
6000	22	male	5'11"	188	7.00	5.33
6001	36	male	5'9"	180	7.17	6.50
6002	28	male	5'9"	180	7.67	7.25
6004	39	male	5'11"	205	7.33	5.67
6005	21	male	5'6"	160	6.92	4.50
6006	30	male	5'8"	245	7.58	8.00
6007	32	male	5'11"	170	7.25	5.50
6008	24	male	5'11"	187	7.08	5.00
6009	32	male	6'1"	215	6.83	4.75
6011	34	male	6'0"	190	7.00	5.67
6014	22	male	5'6"	135	7.58	4.75
6017	25	male	5'4"	130	7.17	6.25
6019	28	male	6'0"	170	6.67	7.00
6020	24	male	5'5"	125	8.08	8.00

possible volunteers for participation in the experiment. Interested candidates were asked to call the Unit for Experimental Psychiatry at the University of Pennsylvania School of Medicine to receive more details of the experimental protocol from a knowledgeable research investigator. During this initial phone conversation, individuals received a preliminary phone screening to evaluate the potential of inclusion for the experiment. Questions asked were derived from sleep and medical history interview forms that have been successfully used in other sleep deprivation studies. Permission to interview the subject was solicited prior to questioning. All responses to all questions were treated confidentially. Those volunteers who reported no history of medical or sleep problems and who had the full protocol explained to them over the phone were asked to schedule a laboratory screening if they wished to pursue participation in the experiment.

PROCEDURES

Pre-experimental Screening

Approximately l-2 weeks prior to the experiment, potential subjects reported to the laboratory for a 2-hour laboratory screening session and were provided with exact details of the protocol and they gave fully informed consent. The session also included a confidential medical screen, administered by an investigator, consisting of a question/answer period regarding their health and medical history, as well as a series of questionnaires about their usual patterns of sleep and experiences with sleep deprivation to ensure that they were healthy (e.g., free of sleep disorders) and had a stable sleep/wake cycle. Subjects were excluded at this stage if they wore corrective lenses (i.e., glasses or contact lenses), or if they could not remain awake for 42 hr without wearing their corrective lenses. Exclusion of persons wearing corrective lenses was necessitated by the need to test the Alertness Monitor of MTI Research, Inc., which required subjects to wear special safety glasses on which the MTI technology was mounted (this is explained further in following sections). There were a number of serious logistical reasons relating to the timeline and costs of the experimental protocol that prohibited fitting the MTI safety glasses with corrective lenses for those subjects who needed them.

Following initial screening, qualified subjects received a wrist actigraph (a miniature ambulatory microprocessor unit that detects and records movement) and a sleep diary, to track their habitual sleep/wake cycles for 1 week. During this week they were also asked to call a voice mailbox immediately prior to going to bed and upon awakening in order to record their sleep times. Determinations of the regularity and normalcy of their sleep/wake profiles were made by integrated evaluation of actigraphic and diary data and call-in times. Only those subjects with healthy, nocturnally-placed sleepwake cycles were permitted to enter the 42-hr

TSD laboratory protocol. Eligible subjects who met all inclusion criteria were scheduled for the laboratory protocol, but continued to wear wrist actigraphs, complete sleep diaries and call-in their bedtimes, as a quality assurance that they had stable sleep/wake cycles up to the time of the experimental protocol.

Experimental Protocol

At the end of 1-2 weeks of ambulatory monitoring, subjects were asked to report to the laboratory at University of Pennsylvania School of Medicine at 7:00 am. to undergo a period of 42 hr of sustained wakefulness and testing. Subjects were studied in pairs, although more subjects could have been studied if more equipment for each technology had been available. Throughout the 42-hr period without sleep, subjects wore a number of fatigue-tracking technologies (see below) that varied in the biobehavioral measures they acquired: three recorded aspects of eye blink/closure; two involved brain wave activity (EEG) algorithms; and one device recorded head movements. (Separate monitoring was also carried out on EEG and eye movements [EOG] to assess physiological sleepiness [i.e., microsleeps], which was not performed as part of the contract and will not be reported here.) EEG and EOG data were acquired using new Oxford Instruments (Clearwater, FL) digital ambulatory recorders (Medilog MR95) and associated digital EEG replay systems (Vision Systems). All data were acquired by the above systems only during the performance testing sessions of the protocol. Circadian phase was also monitored by recording body temperature orally every two hours, at the end of each performance bout.

Throughout all portions of the 42-hr experimental protocol, trained staff members remained with subjects, to ensure wakefulness was maintained; to check and maintain low electrode impedance; to adjust and calibrate equipment and technologies; to facilitate

subjects performing appropriately during test bouts; and to ensure that appropriately timed events occurred according to the protocol. Biobehavioral monitors also engaged in social activity with subjects to assist them in staying awake between performance test bouts, and they unobtrusively monitored subjects during all test bouts to ensure that subjects did not fall asleep and cease to attempt performing. Monitors recorded all pertinent information that occurred throughout all phases of the study, including any difficult times the subject may have experienced in remaining awake. If a subject fell asleep (stopped performing) for a period of 30 sec. during the performance test battery, the monitor alerted the subject by calling the subject's name. If the subject failed to respond, the monitor gently touched the subject on the shoulder until a response was solicited and the subject continued with the test battery. Technical monitors were responsible for the proper use and calibration of all technical equipment of the device/algorithm being tested, as well as making sure that all technologies were functioning properly. If a technology failed, it was the task of the technical monitor to troubleshoot the situation and make proper adjustments to assure the integrity of the collected data. They also had the responsibility of initiating and terminating data collection for technologies during the 20-min. PVT performance task.

Every 2 hr throughout the 42-hr period of sleep deprivation, subjects performed a 1-hr computerized neurobehavioral assessment test battery (NAB). The test battery included a 20-minute reaction time test, a 2-minute memory test, a 2-minute symbol substitution test, a time estimation test, a 5-minute visual fixation test and a 15-minute tracking task. Due to the fatiguing nature of the tracking task, subjects were asked to perform it every 4 hr rather than every 2 hr. During the test bout subjects were also asked to make a number of different psychometric ratings about their mood, alertness, and performance. Throughout the testing portions of the 42-hr period of wakefulness, subjects remained in constant light (< 150 lux) in a room that was temperature-

controlled without time-cues. Between test bouts subjects remained in constant light (< 800 lux) in a room in which they had access to a television and VCR. Meals (excluding caffeine products) were scheduled at regular intervals (breakfast, lunch, supper, late night snack) throughout the 42-hr period. Physical activity was limited to sedentary and relatively passive activities during TSD. Table 2 displays the test bout sequence for the experiment.

NEUROBEIIAVIORAL TEST BOUT

The computerized neurobehavioral assessment battery (NAB) administered once every 2 hr to subjects during the 42-hr TSD included both subjective and objective measures (Dinges & Powell, unpublished as a computerized ensemble). The NAB computer ensemble contains extensive data reduction analysis software that automatically extracts multiple performance and subjective metrics, utilizing appropriate criteria and transformations. The test battery required 40 min. to perform, permitting its use as a repeated neurobehavioral probe throughout the study. The NAB performance battery included the following performance tests, which have been validated to be sensitive to experimentally-induced sleep loss: (1) Psychomotor vigilance task (PVT; see below) yields six highly informative metrics on the capacity for sustained attention and vigilance performance (Dinges & Powell, 1985; Dinges & Kribbs, 1991); (2) a probed recall memory (PRM) test that controls for report bias and evaluates free recall/retention (Dinges et al., 1993); (3) a digit symbol substitution task (DSST) that assesses cognitive throughput (speed and accuracy); and (4) Performance Evaluation and Effort Rating Scales (PEERS) to track self monitoring, compensatory effort, and motivation (Dinges et al., 1992). NAB subjective activation tests included the following: (1) Stanford Sleepiness Scale (SSS) (Hoddes et al., 1973); (2) visual analog scales (VAS) for sleepiness-alertness, and mental and physical

exhaustion; (3) Activation-Deactivation Checklist (AD-ACL) (Thayer, 1986); (4) Karolinska Sleepiness Scale (KSS) (Bikerstedt & Gillberg, 1990); and (5) Profile of Mood States (POMS)

Table 2. Test bout sequence during 42-hr TSD period.

Performance bout number	*Time of day*	*Duration (hr:min)*	*Total hours awake*
1	10:10am - 10.50am	0:40	4
break		0:55	
2	11.45am - 12.40pm	0:55	6
break		1:30	
3	14.10pm - 14.50pm	0:40	8
break		0:55	
4	15.45pm - 16.40pm	0:55	10
break		1:30	
5	18.10pm - 18.50pm	0:40	12
break		0:55	
6	19.45pm - 20.40pm	0:55	14
break		1:30	
7	22.10pm - 22.50pm	0:40	16
break		0:55	
8	23.45pm - 00.40am	0:55	18
break		1:30	
9	02.10am - 02.50am	0:40	20
break		055	
10	03.45am - 04.40am	0:55	22
break		I:30	
11	06. 10am - 06.50am	0:40	24
break		0:55	
12	07.45am - 08.40am	0:55	26
break		I:30	
13	10.10am- 10.50am	0:40	28
break		0:55	
14	11.45am - 12.40pm	0:55	30
break		I:30	
15	14.10pm - 14.50pm	0:40	32
break		0:55	
16	15.45pm - 16.40pm	0:55	34
break		1:30	
17	18.10pm - 18.50pm	0:40	36
break		0:55	
18	19.45pm - 20.40pm	0:55	38
break		1:30	
19	22.10pm - 22.50pm	0:40	40
break		055	
20	23.45pm - 00.40am	1:30	42

(McNair et al., 1971). The eyes-open test required that subjects fixate on a visual cue such as a red dot for a period of 5 minutes. The tests described above were completed by the subject for every bout performed approximately every 2 hr throughout the 42-hr period of waking. Subjects then completed a Compensatory Tracking Task (CTT) for a duration of 15 minutes every other bout. This test is a visual tracking task that requires the subject to keep a moving circle within target with additional forces acting upon it (Makeig & Jung 1996). The CTT was required for calibration of the EEG algorithm developed by Scott Makeig and colleagues. Table 3 summarizes the items administered in the neurobehavioral test bout.

Table 3. Items in neurobehavioral test bout.

Neurobehavioral Assessment Battery	**Time (mins)**
Effort to Stay Awake Rating (ESA)	0.25
Stanford Sleepiness Scale #1 (SSS)	0.50
Visual Analog Scales #1 (VAS)	0.25
Probed Recall Memory, Presentation (PRM) *	0.50
PVT Pre Test Mood Rating (PPMD #1)	0.25
Psychomotor Vigilance Task (PVT) *	20.00
PVT Post Test Mood Rating (PPMD #2)	0.25
Probed Recall Memory, Recall (PRM) *	2.00
Activation-Deactivation Adjective Checklist (AD-ACL)	1.00
Digit-symbol Substitution Task (DSST) *	2.00
VAS #2	0.25
Time Estimation Task *	3.00
sss #2	0.50
Karolinska Sleepiness Scale (KSS)	0.50
Profile of Mood States (POMS)	2.50
Performance Rating	0.25
Effort to Perform (SEQ1)	0.25
Effort Expended (SEQ2)	0.25
Post-Test Alertness Rating	0.50
Eyes Open Test	5.00
Compensatory Tracking Task (CTT) o	15.00
Total Time =	55.00

*Performance task that was performed every bout.
oPerformance task performed on alternate bouts.

PVT LAPSES AS VALIDATION CRITERIA FOR TECHNOLOGIES

The 20-min. PVT task completed by subjects every 2 hr for 42-hr of waking, yielded the PVT lapse performance data that was used as the validation criterion for technologies. The PVT task is a "simple" (as opposed to multiple choice) reaction time (RT) test designed to evaluate the ability to sustain attention and respond in a timely manner to salient signals (Dinges & Powell, 1985). The sensitivity of lapses to drowsiness combined with the performance features that the PVT shares in common with driving (i.e., requirements for sustained attention + rapid responses), make PVT lapses a reasonable selection as a validation criterion for drowsy driving technologies. PVT performance has been demonstrated to be highly sensitive to changes in alertness/drowsiness associated with circadian phase (Dinges & Kribbs, 199 1; Wyatt et al., 1997); with acute total sleep deprivation (Dinges et al., 1994); with cumulative partial sleep loss (Dinges et al., 1997; Rowland et al., 1997); with sleepiness in the elderly (Samuel et al., 1996); with shift work /jet lag (Rosekind et al., 1994); with the demands of medical house staff (Geer et al., 1995; Smith-Coggins et al., and Howard et al., unpublished studies, Stanford University); and with untreated obstructive sleep apnea in clinical populations (Kribbs & Dinges, 1994), and untreated apnea in commercial motor vehicle operators (Dinges et al., in press). The lack of contamination of PVT performance by learning curves and aptitude, and its documented sensitivity to loss of alertness and increasing sleepiness/drowsiness induced by experimental sleep loss, by occupational sleep loss, and by medically-based sleep loss, makes the PVT an excellent performance criterion for testing the validity of on-line, biobehavioral monitors intended to detect performance-impairment from hypovigilance, fatigue, or drowsiness.

The PVT hardware and software were invented by David F. Dinges,.Ph.D., and

implemented by Mr. John W. Powell, IV. The task was designed to be simple to perform, free of a learning curve or influence from acquired skills (aptitude, education), and highly sensitive to an attentional process that is fundamental to normal alert functioning. The PVT and its resulting metrics are based on a model of changes in brain function induced by sleepiness/drowsiness (Dinges, 1989). For the current study, the PVT task was incorporated into the PC on which all other performance tests were also performed. The task consisted of responding to a small, bright yellow light stimulus on the computer screen by pressing a response button as soon as the stimulus appears, which stopped the stimulus counter and displayed the RT in milliseconds for a I-sec. period. The inter-stimulus interval varied randomly from 2 sec. to 10 sec., and the task duration was 20 mins. The subject was instructed to press the button as soon as each stimulus appeared, in order to keep the RT number as low as possible, but not to press the button too soon (which yielded a false start [FS] warning on the display). At the beginning and end of the PVT task a sleepiness visual analog scale was presented.

The frequency of PVT lapses (i.e., RTs > 500ms), as automatically extracted from PVT files by computer, were selected as the primary criterion variable, due to their well-documented sensitivity to sleepiness (Dinges & Kribbs, 1991; Dinges, 1992; Dinges et al., 1994;Kribbs & Dinges, 1994; Dinges et al., 1997). However, since a simple lapse frequency metric does not incorporate total time in the lapse state, all data were also validated against a second PVT lapse criterion, namely cumulative lapse duration time. A summary of the results of this second extensive validation analysis are contained in an Appendix. While there were small differences in coherence for comparisons of PVT lapse frequency criterion to PVT lapse duration criterion, The basic findings were the same (see Appendix).

Lapses reflect the most serious loss of attentional capability, because they represent a

failure to respond (or a failure to respond in a timely manner) to a signal the observer is monitoring. As described in the study design section (see above), in addition to the validation criterion variable of total PVT lapses for each 20-min. PVT test (i.e. global index of vigilance performance impairment for that bout), the number of lapses in each 1-min. of PVT performance throughout each 20-min. PVT test were also segregated (i.e. minute-to-minute fluctuations in vigilance performance) (Kribbs & Dinges, 1994).

Following data acquisition at the University of Pennsylvania, each technology supplier received the raw data for their device/algorithm gathered in 20 discrete files (i.e. 20 files for each unique subject). As per the double-blind procedures summarized in the study design section (see above), the files were coded in such a way that there was no information regarding the time of the performance bout in which the information was acquired, except for test bout 1. In other words, the random sequences of files provided to each supplier for their device/algorithm was the same random sequence as the data given to other participants, but it did not convey how long a subject had been awake, time of day, or the subject's PVT performance during the bout. Upon receipt of their data, each supplier was asked to calculate a global drowsiness score for each 20-min. PVT test bout, and a l-min. drowsiness score for each of the 20 min. in each of the 20 test bouts. This yielded a total of 420 drowsiness scores, 400 one-minute scores (for calculating minute-to-minute coherence), and 20 global (20-minute) scores (for calculating bout-to-bout coherence), per subject per technology/algorithm. Suppliers were then required to send the 420 drowsiness scores for each subject to the University of Pennsylvania investigators for statistical analyses in relation to PVT lapses -- being careful to ensure that each drowsiness index was identified from the file in which it was derived and the minute in which it occurred.

TECHNOLOGIES

A total of six technologies were selected for study based on a combination of criteria. However, two of the technologies had more than one drowsiness metric, which resulted in a total ‘of nine drowsiness metrics being studied. (Subsequent analyses in the “Results” section revealed high intercorrelations among the drowsiness metrics within each of these two technologies). In terms of selection of the 6 technologies for study, some of the eye blink/closure technologies and EEG algorithms were of interest to FHWA--OMC, and some were of interest to NHTSA, based on previous research or on potential utility in motor vehicle operation. Other technologies were included.by the project Principal Investigator (David F. Dinges, Ph.D.) for comparison purposes. Inclusion of a technology/algorithm in the study was not an indication of endorsement by either the federal agencies supporting the project or by the universities performing the research. Similarly, technologies/algorithms not included in the project should not be construed as a sign that other technologies lack potential for alertness/drowsiness detection. Table 4 provides a summary of the source of the technologies/algorithms included in the validation study.

Eye/Facial Ratings

PERCLOS (P70, P80, EM)**.** A low-light, cIosed-circuit television camera was used to monitor the subject’s entire face for recording eyes and eyelids during PVT performance testing, as a data acquisition system for the PERCLOS measures (Wierwille et al., 1994). According to a study performed by Wierwille et al. (1994), drivers in an automobile simulator exhibit certain characteristics when fatigued, that can be easily observed in eye and facial changes (Wierwille & Ellsworth, 1994). Alert drivers were reported to have normal facial tone, and fast eye blinks with short ordinary glances. Drowsy drivers were reported to have decreased facial tone, slower eyelid

Table *4:* Summary of technologies/algorithms evaluated.

Supplier name	*Corrtact person*	*Phone number*	*Device/Algorithm*	*# ss studied*	*Method to obtain*
eye/facial ratings					
Carnegie Mellon Research Instit. PO Box 2950 700 Technology Drive Pittsburgh, PA 15230-2950	James J. Staszewski, Ph.D. Richard Grace, Ph.D.	4 12-268-888 1 4 12-268-3493	PERCLOS (Wierwille et. al., 1994) measure of the proportion of time that the eyes are > 80% closed over a one minute interval	I0	video of face
EEG algorithms					
Consolidated Research, Inc. 26250 Euclid Avenue, Suite 24 Euclid, OH 44 132	Richard Kaplan, Ph.D. President	2 16-289-233 1	**EEG algorithm**	4	EEG electrodes at 02 referenced to A l A2
Naval Health Research Center PO Box 85122 San Diego, CA 92 186-5 I22	Scott Makeig, Ph.D.	619-553-8416	**EEG algorithm** adjusted by Compensatory Tracking Task data (Makeig & Jung, 1996)	4	EEG electrodes at C4, O l, F3, P4 referenced to AlA 2; EOG electrodes at LOC-ROC
head motiorr sensor metrics					
Advanced Safety Concepts, Inc. PO Box 2534 Sante Fe, NM 87504	Philip Kithil President	505-984-0273	**Proximity Array Sensing System** (PASS) head position monitor / metric .	5	I" x l5" x 18" array of overhead capacitive detectors
eye blink monitors					
MTI Research Inc. 7 Littleton Road Westford, Ma 01886	Ed MacLeod President	978-692-9898	**Alertness Monitor** ambulatory eye blink monitor	14	infrared emitter/detector mounted on safety glasses
IM Systems, Inc. 1055 Taylor Avenue, Suite 300 Baltimore, MD 2 1286	David Krausman, Ph.D. Vice-President	4 l O-296-7723	**Blinkometer** ambulatory eye blink monitor	6	sensor placed on outer canthus of the eye

movements and longer eyelid closures (2+ sec.) accompanied by eye movements that rolled upward and sideways. Wierwille and colleagues (1994) developed a metric of drowsiness referred to as "PERCLOS," based on the above observations. PERCLOS is a measure of the proportion of time that the eyes of a subject are closed over a I-minute period as judged by a human scorer (from videotapes of the subject's face). According to Wierwille and colleagues (1994), PERCLOS heavily reflects slow eyelid closures (rather than blinks), which can be construed as both a physiological indicator of drowsiness, as well as an indicator of interruption in visual information gathering. By selecting certain cut-offs for the proportion of time a human scorer judged the eyes to be closed, drowsiness metrics are extracted from PERCLOS. Hence Wierwille and colleagues (1994) have established a drowsiness criterion of the. eyes judged to be > 80% closed. The resulting metric is the proportion of time in a minute that the eyes met the 80% closure criterion.

The video camera systems used to record eye/facial PERCLOS ratings included a Panasonic Color CCTV Camera, a JVC Color Video Monitor and a Mitsubishi Hi-Fi Video Cassette Recorder. The Panasonic color camera (model WV-CP220) had an electronic light control function and used fl.2 aperture lenses. The camera was positioned approximately 45^{o} to the right of the subject at approximately chin level and angled upward to gain the best possible full image when the subject's eyes began to close. It was also zoomed to a position that allowed the subject's face to occupy the maximum area of the connected viewable video monitor. To ensure that a subject's face remained properly positioned in the camera's field of view throughout each of the neurobehavioral performance test bouts, video signals were also sent to a JVC Color Video Monitor (model TM-I 3 1 SU), which was continuously observed by a staff member during data acquisition. This also allowed the biobehavioral monitor to view the

subject's face and eyes to confirm that the subject remained awake throughout performance testing trials. The color video images of each subject's face were also recorded, throughout each 20-min. PVT bout, on a Mitsubishi Hi-Fi Video Cassette Recorder (model HS-U560), which provided a high resolution VHS color picture that was used for scoring of PERCLOS measures by trained scorers at Carnegie Mellon Research Institute (CMRI). Before being sent to CMRI for PERCLOS scoring, the video tapes were copied with a scrambled test bout sequence (see Study Design section above) in order to keep CMRI scorers blind to the exact time of day or the number of hours each subject had been awake. No audio information was available to PERCLOS coders.

Video recordings were collected on all 14 subjects who participated in the protocol but video recordings for only 10 subjects were storable by CMRI. The unscorable video recordings were due to the fact that subjects were wearing an updated version of the Alertness Monitor (MTI Research, Inc.) glasses that contained plastic lenses that produced a glare from the computer screen. Reflection was not an issue for the first 10 subjects since the earlier models of the Alertness Monitor did not contain lenses.

The PERCLOS scoring procedures used by CMRI followed those of Wierwille et al. (1 994). Trained scorers used a linear potentiometer to continuously and manually track any movement of the subject's eyelids. Ratings of PERCLOS values were accomplished as follows. At the beginning of the first training session, all potential coders were given a brief introduction to the background and purpose of the study by CMRI investigators. They were told that the project's aim was to develop an early detection system for truck-driver drowsiness, and that behavioral measures such as percentage eye closure were going to be used in the development of a reliable indicator of driver drowsiness. Percentage of pupil/iris coverage was stressed as the

operational definition on which eye closure would be based. Coders were provided with a computer-generated eye at 0%, 25%, 50%, 75%, and 100% closure (see Figure I) at all times during the scoring of PERCLOS variables. All coders had eye/facial scoring training before the actual scoring of videos from the 42-hr TSD protocol--they had approximately 6 hr training on simulator tapes in addition to 2 hr training on video tapes of both alert and drowsy truck drivers taken from field studies performed by CMRI. Coders were given the following specific scoring instructions.

The purpose of this study is to investigate the relationship between changing physiological states and human performance on complex tasks. It focuses on the relationship between eye closure and individuals' performance on a simulated driving task. Your job will be to judge, as accurately as possible, the degree to which a person !s eyes are closed from moment to moment as (s)he performs this task.

To rate eye closure, you will be seated in front of a television monitor. 0n it videotaped segments will be presented showing a driver's face as (s)he controls a simulated car. You are to continuously rate the degree to which the driver's eyes are closed by moving the sliding arm on the apparatus, which will be placed to the side of your preferred hand When the siider is up as far as it will go, it corresponds to the eyes being 100% open; likewise, when the slider is down as far as it will go, it corresponds to the eyes being 100% closed. You should focus on the driver's eyes, continuously judging rheir degree of closure and moving the slider proportionately to record your judgments. The attachedfigure provides a standard to help you judge what different degrees of eye closure Iook like.

Because paying attention both to the driver's eyes and to the position of the slider simultaneously is difficult, a lighted display attached to the side of the monitor will provide feedback on the current position of the slider. It is there to enable you to monitor the position of the siider without turning your eyes away from the video screen. Each Iighted element in this display represents approximately 3% of closure.

Moving the slider down from the top causes successive elements to light up. The color of the lights will change as the slide moves past certain levels: The color changes from green to orange (or vice versa) as the slider moves past the 50% level. As the slider moves past the 75% level, the lights change from orange to red.

With the exception of fast blinks (that is, blinks occurring in less than l/4 sec.). we want you to register all changes in eye closure. Since blink duration is a highly subjective measure, expect that it will take some practice to be able to successfully discriminate which blinks are fast from those that are not. As you are rating, also expect to feel; some lag time between the actual eyelid movement and your reaction.

This is acceptable, as long as you try to keep the lag time constant. In other words, don' t feel as if you have to react more rapidiy to a blink than to a slow eye movement.

At certain times in your rating, you may not be able to fully see both eyes If this is the case, rate according to whatever portion is visible- At other times, you might notice that both eyes don't match in their level of closure. If this is the case, rate according to the eye that is most open (i.e. the dominant eye). Finally. the person whom you are rating may at some point tilt his/her head back in the chair.

This will cause their eyes to appear more closed than they actually are, so remember to always rate according to the percentage of pupil / iris coverage. At all times, try to make accuracy of measurement your first priori@

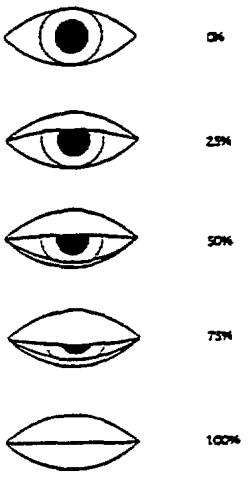

Figure 1. Schematic of computer-generated eye closure at 0%, 25%, 50%, 75%, and 100% closure provided to CMRI coders of PERCLOS variables.

Two coders were selected to score all video segments collected. Scoring was done independently by each coder on a 19" color television monitor. Coder 2 scored subjects 6000, 6001,6002,6004,6009 and 6011 (segments 7-20), while coder 1 scored subjects 6005,6006, 6007,6008 and 6011 (segments l-6). Inter-rater reliability was calculated on three separate 20-min. records (subject 6006, segment 3; subject 6005, segment 2; subject 6009, segment 4) for each of the three PERCLOS variables (P70, P80, eye measure). P70 was the proportion of time the eyes were judged to be 70% to 100% closed. P80 was the proportion of time the eyes were judged to be 80% to 100% closed. EYEMEAS (EM) was the mean square percentage of the eyelid closure rating. Inter-rater reliabilities for the single subject segment available were

r = 0.91 for P70, r = 0.91 for P80, and r = 0.95 for eye measure. Intra-rater reliabilities for the single segments from two subjects on which it was available were r = 0.77 and r = 0.99 for P70; r = 0.78 and r = 0.99 for P80; and r = 0.87 and r = 0.99 for eye measure.

To avoid misunderstanding the instructions, coders were also given verbal instructions to keep the lag constant for all scorable eye closures, and to ignore fast eye'blinks. After a tape was coded completely, a research supervisor would go through the tape from beginning to end, marking the time and duration of all instances when the driver's pupils were out of view of the camera Out-of-view segments were excluded with a l-second delay to account for coder reaction time. All coder data associated with these out-of-view segments were then excluded from analysis. In order for a frame to be considered "out-of-view," it had to be the case that the driver moved his/her head out of the viewable range of the camera such that neither pupil could be seen.

EEG Algorithms

Electroencephalographic (EEG) methods capable of detecting changes in brain wave activity associated with fatigue are commonly used in the assessment of fatigue and sleep onset. Studies have shown that changes in EEG activity of the brain do reflect changes in alertness levels and have potential as a basis for the detection and management of fatigue (Makeig et al., 1993). However, the range of variables that might influence the utility of EEG as an alertness/drowsiness detection system are quite large (e.g., electrode number and location; EEG frequencies and/or amplitudes analyzed, and other specific aspects of EEG signal processing). The protocol included two different EEG monitoring algorithms.

Consolidated Research Inc. (CRI) EEG algorithm. One of the EEG algorithms tested was CRI' s Drowsiness Detection Algorithm, which is described as using "specific identified EEG waveforms" recorded at a single occipital site (01 or 02). The algorithm is reported by CRI

to be capable of continuously tracking an individual's alertness and/or drowsiness state through alert periods, sleep periods and fatigued periods as well any changes in alertness levels. The algorithm uses approximately 2.4 sec. of EEG data to produce a single output point with a 1.2 sec. update rate. The algorithm output is an amplitude variation over time that increases in magnitude in response to the subject moving from normal alertness through sleep onset and the various stages of sleep. Because of the fast update rate, the algorithm is sensitive to transient changes in alertness on a second-by-second basis. Therefore, the supplier reports that "microsleeps," "microarousals," and other transitory phenomena are easily distinguishable in the output.

Unlike some EEG algorithms for alertness/drowsiness detection, CRI's algorithm for predicting drowsiness state does not rely on electrooculographic (EOG), or any other measurement of eye movements or the status of the eyes. CRI has reported that episodes where subjects closed their eyes due to increased sleep pressure (unintentional closure) were typically accompanied by large deflections in their EEG algorithm output measure. Therefore CRI asserts that their EEG measure is tracking a state internal to the subject that is related to excessive drowsiness, sleep pressure or sleep. In contrast, CRI notes that their algorithm output is not a direct measure of reactions or reaction time. They anecdotally report that during performance testing intentional eye closures not caused by acute sleep pressure can result in missed signals, but not always distinguishable in the output measure from their algorithm. CRI maintains therefore that their EEG output measure is only an indirect measure of task performance, and that the latter can be subject to variation from other sources such as individual differences, motivation, experience with sleep loss, etc.

For the current study, EEG data for CRI's algorithm was collected on only 4 subjects due to the fact that the CRI equipment specified and provided by the company for data acquisition was available for a limited period of time. Moreover, the Oxford Instruments Medilog equipment used for Scott Makeig's EEG algorithm (described in the next section) was not used for CRI's algorithm. The CRI EEG recording site was O2 referenced to A1A2, and data were acquired through a Grass PolyVIEW preamplification system and Model 15 Neurodata Acquisition System. The CRI system began logging data upon receipt of the first stimulus-response (S-R) event signal output by the NAB computer on its parallel port. This signal and the false start event signal were buffered by optical-isolator circuits before going to the CRI amplifiers and the MR95 reorder. CRI's EEG data were collected on their system's hard disk, then later transferred to "JAZZ" disks and sent to CRI Research for off-line analysis. CRI provided a "mean" measure of alertness/drowsiness for bout-to-bout coherence assessments and a "mean" measure for minute-to-minute coherence assessments. These "mean" measures could be applied across all subjects and they therefore were used in the coherence analyses. CRI also provided four "alternate" minute-to-minute measures of alertness/drowsiness that were idiosyncratic to each subject on which data were available. These alternate measures were not analyzed for this report.

Dr. Scott Makeig's (SMM) EEG algorithm. The second EEG algorithm tested was developed by Scott Makeig, Ph.D. and colleagues at the Naval Health Research Center, San Diego and Salk Institute, La Jolla, California, and was based on methods for modeling the statistical relationship between changes in the EEG power spectrum and changes in performance caused by drowsiness. The algorithm is reported to be a method for acquiring a baseline alertness level, specific to an individual, to predict subsequent alertness and performance Ievels for that

person. Baseline data for preparing the idiosyncratic algorithm were collected from each subject while performing the CTT (see Table 3).

The CTT task requires subjects to keep a moving circle within a certain radius of a ring, both displayed on a computer screen. The circle is constantly moving due to three different forces acting on it. The forces, aside from the user input force, prevent the circle from being within a certain distance of the ring for more than 95% of the time when there is no user input. Due to the heavy performance burden the CTT placed on subjects in combination with the other performance tests, the CTT was only administered in every other performance bout throughout the 42-hr period of wakefulness. Consequently, unlike all other technologies/algorithms tested in the study. the SMM EEG algorithm was only available at a temporal resolution of once every 4 hr. rather than once every 2 hr.

Makeig and Inlow (1993) have reported drowsiness-related performance decrements occurring in irregular cycles of 4 min. and longer. They have attempted to estimate performance fluctuations from EEG spectral changes integrated over shorter time windows of 30-90 sec. However, they have observed that an individualized EEG model for each subject is essential due to large individual differences in patterns of alertness-related change in the EEG spectrum (Makeig & Inlow, 1993; Jung, et al., 1997). The method of EEG-based alertness monitoring they developed involves several signal processing stages (Jung et al., 1997). EEG data recorded from l-8 scalp channels are examined for excessive muscle activity, and noisy epochs are rejected. Eye movements are regressed from the scalp data using electrooculographic (EOG) reference channels. Next, the scalp data are converted to log spectra using Fast Fourier transforms (FFTs) and median-smoothed to ignore outliers. Final smoothing is performed using a 30-90 sec. moving window. CTT performance data, collected at an average rate of at least 10 measurements

per second (17Hz), is smoothed with the same moving window, taking care to align the leading edge of successive performance windows with the leading edge of the corresponding EEG windows.

After data smoothing, EEG spectral data in the range OS-25 Hz are transformed using singular value decomposition into the eigenvector basis. Only the first 4-5 eigendimensions contain information about performance (Makeig & Jung, 1995). The remaining eigendimensions are therefore discarded, and the first 4-5 are input into multiple linear regression or backpropagation neural networks that learn, deterministically or stochastically, to estimate concurrent performance changes solely from the changes in the EEG spectrum. Once trained, these linear or nonlinear network models can then be applied to smoothed EEG spectral data from subsequent task periods on the same subject, and appear to deliver accurate moving performance estimates in near-real time (Jung et al., 1997).

EEG data for SMM's algorithm were acquired from 4 subjects at scalp electrode sites C,, 0,, F,, P4 referenced to linked mastoids (A1A2), using the new, digital, palm-sized, ambulatory Oxford Medilog Recorder (MR95) belonging to the P.I. EOG data were also acquired from LOC-ROC. The MR95 was powered by rechargeable batteries and worn by subjects throughout all test bouts. EEG, EOG and related physiological data were stored digitally and internally on miniature hard disk (Oxford Instruments). Impedance measurements were performed every 6 hr throughout the 42-hr protocol to maintain impedance below 5 k o for all EEG and EOG sites. EEG data was collected throughout the entire neurobehavioral test battery, but was event-marked only during the 20-min. PVT test in every bout. A PVT stimulus response (S-R) channel and false start (FS) channel were designated on the MR95 to allow for time locking with EEG. This allowed researchers to observe, in exact time, EEG changes that occurred with the presentation

of a stimulus and the time required for the subject to respond. EEG and EOG data for Dr. Makeig were downloaded from the MR95 to the new Vision Replay System (Oxford Instruments), and then converted to ASCII files for transfer to Dr. Makeig.

Further time-locking also occurred between video recording and collected EEG data with the use of Oxford's Video Interface Processor (VIP). The VIP allowed for the synchronizing of video tapes with EEG data. The VIP stamps a time and date code sequence on the video recordings that permits precise analysis of performance lapses in conjunction with EEG recordings. These analyses are being developed by the **P.I.,** however, and do not pertain to the current report.

Head Position Monitoring Device

Head position is believed to change with increasing levels of fatigue. With increasing fatigue, a person may begin to lose muscle tone in the neck and the head may begin to bob, drop or roll, which can be characteristic signs of sleepiness. It has also been speculated that control of head motions may change depending on the degree of alertness of an individual. A device developed to detect fatigue based on head motion was tested in the current study.

Advanced Safety Concepts (ASC). Inc. Proximity Array Sensing System. The non-contact Proximity Array Sensing System (PASS), developed by Advanced Safety Concepts, Inc. is an apparatus designed to record the x, y and z coordinates of the head at electronic rates using three electromagnetic fields. Its development is based on research that indicates a relationship between micro-motion of the head and impairment or drowsiness. It is hypothesized by ASC that changes in the x,y,z coordinates of the head may be an indicator of fatigue onset, and that PASS may detect micro-sleeps based on different head movement patterns. Advanced Safety Concepts,

Inc. reports that in laboratory tests, the PASS system has detected changes in head position as little as 0.0 l", while providing absolute XYZ resolution of head position to about 0.1."

The ASC, Inc. PASS system studied employed an array of three capacitive sensors that create hemispheric sensing fields encompassing a seated person's head position. The detectors were contained in a foam-core module (1" x 15" x 18") mounted above the subject's head, and connected to an electronics module (2.5" x 10" x 12") provided by ASC, Inc. The PASS triangulated the position of the head by determining the proximity of the head to each capacitive sensor through partial blocking of the sensing fields. The intersection of the three proximities defined a point in space that equated to the center of the head. This point was tracked over time to determine patterns of head motion. The detected voltages were transmitted to a computer that contained the processing algorithm developed by ASC, Inc.

PASS data were collected on 5 of 14 subjects who participated in the 42-hr TSD protocol. The limited data collection was due to the fact that ASC, Inc. was included in the study after-two protocols had already been completed, and that they were only able to provide one PASS system to be used in each of the scheduled protocols, making it possible to collect data on only 1 of 2 subjects studied in each protocol session. The PASS apparatus provided was positioned rostral to and as close to a subject's head as possible to allow for accurate data collection and was realigned before each performance testing bout began. The PASS head position detectors were connected to a laptop computer with Intel 80486 cpu for data collection and storage. Data collection during the PVT performance test was initiated and terminated manually on the laptop by a technical monitor. Each subject was continuously monitored when the PASS was collecting data to assure they remained within the range of the detectors. The data acquired by PASS was

sent to Advanced Safety Concepts, Inc. for calculation of two alertness / drowsiness metrics (ASC60, ASC90) for each subject.

Eve Blink Monitors

Eye blink activity has also been studied for its utility to predict fatigue and sleepiness (Stem, 1994). Eye blinks can be influenced by psychological and behavioral variables, and provide information about perceptual and cognitive activity. Different aspects of eye blink amplitude, latency, and rate have been of interest in the assessment of fatigue. Two different eye blink monitors were tested in the current protocol. They varied in both the means of collecting data and the associated algorithms to predict fatigue levels.

MTI Research (MTI), Inc. Alertness Monitor. MTI Research has developed "Alertness Monitor," an eye blink device designed to detect and track fatigue. The Alertness Monitor determines alertness/drowsiness levels by measuring the ratio of eyelid closure to eyelid open, using optical electronics that can be adapted to any style of eye glass frame. For the current study, only subjects who did not wear corrective lenses (i.e., glasses or contact lenses), or those who did but who could remain awake for 42 hr without wearing their corrective lenses, were included. To test MTI's Alertness Monitor all 14 subjects wore special safety glasses on which the MTI technology was mounted (the first 10 subjects had no lenses in their safety glasses to avoid problems with computer screen glare; while the final 4 subjects had plane lenses in the safety glasses). In all cases, the safety glasses had unobtrusive optical electronics mounted in such a position that an emitted infrared beam fell along the axis of the eye blink (especially the eyelid), such that the eyelash could break the beam during an eye blink. The source of the infrared beam was transmitted from an emitter on the nose-piece of the eye glasses to a sensor, located on the arm of the eye glasses. Both the emitter and sensor were mounted on the eye glass

frames in a way that minimized the risk that the infrared beam would shine into the subject's eye. If the beam did shine into the subject's eye, the levels of infrared light it emitted were below levels considered hazardous by the standards of the American Congress of Governmental and Industrial Hygienists (ACGIH Standards, 1995 edition).

The research model of Alertness Monitor was powered by a 9-volt battery, allowing data collection for a period up to 24 hr. The Alertness Monitor glasses were not self-calibrating and required that each subject, facilitated by a technical monitor, properly fit the glasses before accurate data collection could proceed. Proper alignment of the glasses was done while Alertness Monitor was switched to a calibration mode that sounded a series of beeps once proper alignment of the glasses had been achieved.

Although the Alertness Monitor glasses were used on all 14 subjects run in the 42-hr TSD protocol, as the study progressed, different models of the Alertness Monitor were being developed by MTI Research, Inc. This ongoing development allowed researchers to study a total of three different Alertness Monitor models (model 1 was tested on the first 4 subjects; model 2 on next 6 subjects; and model 3 [most recent] was tested on the final 4 subjects). More recent models (e.g., model 3) included both different hardware/structural design as well as updated drowsiness algorithms, and were intended to be an improvement over earlier models (e.g., model 1). Eye blink data were collected using a program written by John W. Powell based on information supplied by MTI Research, Inc. The glasses were connected to a nearby PC via a cable and serial port that allowed for data collection and transfer of files to MTI Research, Inc. for application of their algorithm. Technical monitors started and stopped data collection with the beginning and ending of each 20-min. PVT session. However, even though data. collection was stopped after the PVT session, subjectscontinued to wear the Alertness Monitor glasses to keep

the testing conditions constant. MTI Research, Inc. provided a single metric (MTI) of alertness/drowsiness from their algorithm.

IM Systems (IM), Inc. Blinkometer. The "Blinkometer," developed by IM Systems, Inc. is an ambulatory blink recording device that uses an algorithm that is reported by the company to be capable of detecting drowsiness/sleepiness. The Blinkometer can be set on one of two modes: blinks per minute (usually ~20/min) or the blink-to-blink interval. The blinks per minute mode was selected as the data collection mode in the current experimental protocol because it yielded information that was not provided by any of the other technologies tested.

The Blinkometer consisted of a sensor that is placed using a double-sided adhesive disk at the outer canthus of one eye, and a small recording device powered by lithium batteries that could either be placed in a pocket or clipped on a belt. Eye lid activity moved the piezoelectric film in the sensor and an eye blink was detected, which also resulted in a flash from a red LED on the recording unit. The recording device displayed the collected eye blink data on an LCD. The device was calibrated for sensitivity for each individual subject, using a threshold adjustment for blink detection. Subjects were instructed to plug in the Blinkometer sensor to the recording unit and press the event marker prior to each 20-min. PVT, and to again press the event marker and unplug the Blinkometer sensor upon completion of each PVT. A technical monitor oversaw this procedure to assist the subject and confirm that data collection was properly initiated and terminated. Eye blink data were downloaded from the Blinkometer recording unit to a PC, and complete files were sent to IM Systems, Inc. for analysis.

Although the Blinkometer was used on all 14 subjects studied in the protocol, data were lost on a total of 8 subjects. These losses occurred after data acquisition, during downloading of the data through the IM Systems, Inc. interface unit and when IM Systems, Inc. attempted to

segregate data files. The reasons for the data losses remain unknown, but the technical problems with IM Systems' interface and file management components resulted in complete Blinkometer data on only 6 of the 14 subjects on which the Blinkometer was deployed.

IM Systems, Inc. reported that the Blinkometer detects a decreased alertness within 20-30 sec., implementing a fairly straightforward algorithm that the fewer the number of blinks, the greater the level of drowsiness. Data from the baseline session (bout 1) was used to calculate the mean blink rate (number of blinks per minute) for each individual subject--the bout 1 calculated values were considered to be the mean blink rate at full awakening. In subsequent testing sessions, the number of blinks per minute was assigned a drowsiness rating by comparing the blink rate for that minute to the baseline blink rate. IM Systems, Inc. reported that drowsiness scores were assigned as follows: If the number of blinks in an epoch (minute) was less than one standard deviation below the baseline mean or greater, the assigned drowsiness rating was 0 (fully awake). If the number was between one and two standard deviations below the mean blink rate, the drowsiness rating was 1, and so on. The algorithm compared the current blink rate to baseline rate and estimated the number of standard deviations below blink rate--down to a drowsiness level of 5 (between 5 and 6 standard deviations below the mean), providing a single drowsiness score (IM).

STATISTICAL APPROACH

The statistical methodology used to analyze the experimental data was designed to meet several objectives. The primary objective was to enable measurement of the ability of each technology/algorithm to reproduce the relative alertness/drowsiness orderings of discrete time periods as determined by the validation criterion (i.e., PVT lapse frequency and PVT lapse duration [see Appendix]). These were calculated within each subject, across the 20 discrete, 20-

min. bouts collected over the 42-hr sleep deprivation period. To emphasize the fact that two sets of time series were being compared, the term "coherence" from the time series literature (see for example, Brillinger, 1981; Bloomfield, 1976) was used to describe the measures. However, by design, the time series sampling was intermittent (20 min. every 2 hr) and of (relatively) short duration (n = 20 total bouts). Furthermore, it was believed that it could not be ruled out that the bivariate time series contained non-stationary elements arising from subject-specific interactions between circadian rhythms in alertness and natural variation in subjects' abilities to overcome drowsiness brought on by sleep deprivation. This interaction was suggested by the appearance of a step function in PVT lapses over time in some but not all subjects when it appeared that a subject's ability to perform drastically diminished. Thus, although traditionally, the term "coherence" has been used to measure the degree of similarity between two time series in the frequency domain, we computed simple measures of coherence in the time domain using Pearson and Spearman rank correlations for each set of n = 20 bivariate PVT lapse - technology/algorithm pairs. Spearman rank correlations were computed to assess whether Pearson correlation estimates of "coherence" were robust--that is, whether their values were highly influenced by a few extreme values. With rare exceptions, it was found that subject-specific Pearson and Spearman values were consistently similar. Thus, the primary analyses were based on Pearson correlation coefficients referred to hereafter as "coherence" (with PVT lapses). (The meaning of the term "coherence" according to Webster's Unabridged Dictionary, is systematic or logical connection, consistency or dependence proceeding from the natural relation of parts or things to each other. Terms such as "concordance," which Webster's defines merely as "agreement" do not capture the "systematic or logical consistency" across time, as implied in

the term coherence. We believe, therefore, that "coherence" most closely captures the thrust of the analytic approach used in this experiment.)

The resulting coherence values were summarized to determine "average coherence" for each technology/algorithm. In addition, the distributions of coherence values were examined in order to compare among subjects. This was done to determine the extent to which each technology/algorithm yielded consistent degrees of coherence across different individuals. Specifically, interests focused on assessing whether coherence was especially large or small in specific individuals. A secondary objective was to assess whether coherence depended upon subject-specific performance ability. To investigate this, grand means for PVT lapses and for each technology/algorithm were obtained as well as estimates of the between-subject standard deviations. From these values, subject-specific "z-scores" were computed as the average for a specific subject minus the grand mean for all subjects, and the result divided by the between-subject standard deviation. This permitted, for example, a determination of a subject's overall tendency to lapse within the context of this experimental paradigm. PVT lapse "z" values less than zero indicated less than average tendencies to lapse, while PVT lapse "z" values greater than zero indicated more than average tendencies to lapse. Two-sample t-tests were used to compare technology/algorithm coherence values between "higher lapsers" ($z > 0$) and "lower lapsers" ($z < 0$). Similarly, Pearson correlations between PVT lapse z-scores and coherence values were computed to assess the magnitude of the association between coherence and tendency to lapse.

In addition to bout-to-bout coherence, analyses were repeated based on minute-to-minute coherence (i.e., n = 20 bouts times 20 minutes per bout = 400 minutes). This represented a more stringent criterion for coherence than did bout-to-bout coherence. Thus, it was possible that a technology/algorithm could have relatively large bout-to-bout coherence yet fail to as reliably

track PVT lapses in the minute-to-minute time domain. Paired t-tests were used to assess that statistical significance of the mean declines. The difference (decline) in coherence between minute-to-minute values and bout-to-bout values represented a measure of the degree to which drowsiness signals diverged over much shorter time intervals.

A number of tertiary analyses were also performed. (1) As an evaluation of whether coherence changed as a function of overall level of sleep deprivation, coherence values were computed using only performance test bouts 1 through 10 (i.e., 2 - 22 hr awake) and compared by paired t-test to coherence values computed using only performance bouts 11 - 20 (i.e.) 22 - 42 hr awake). (2) To aid in the interpretation of the coherence magnitudes obtained for each technology/algorithm, paired t-tests were used to compare analogous coherence values between PVT lapses and each subject's own ratings of sleepiness based on a visual analogue scale completed at the end of each PVT test. (3) The relationship of coherence results among technologies/algorithms were assessed by way of Pearson correlation coefficients. These inter-technology relationships provided information on whether or not technologies/algorithms were consistently yielded high (or low) coherence for the same subjects. (4) Finally, positive and negative predictive values, and sensitivity and specificity were also computed, over bouts for PERCLOS, the one technology/algorithm that consistently yielded high coherence values, in order to assess on an absolute basis, its ability to correctly signal vigilance lapses.

RESULTS

EFFECTIVENESS OF EXPERIMENTAL DESIGN TO INDUCE PVT LAPSING

To evaluate the validity of the technologies for detecting performance lapse frequency, an experimental design was used in which a wide range of alertness levels was induced by requiring subjects to perform the 20-min. psychomotor vigilance task every 2 hr across a 42-hr period of sustained wakefulness (Le., sleep deprivation). This approach was highly effective. The resulting average number of PVT lapses per performance test bout for the 14 subjects studied are shown in Figure 2. As expected, the mean number of PVT lapses increased by an order of magnitude during the 42-hr test trial ($F_{19,247} = 23.72$, $p = 0.00001$). The resulting bout-to-bout lapse profile shown in Figure 2 reflects the combined effects of the increasing homeostatic drive for sleep and its interaction with the endogenous circadian pacemaker (Dinges & Kribbs, 1991; Dijk et al., 1992; Babkoff et al., 1991). Mean oral temperature readings taken on subjects at the end of each performance bout. controlling for posture, activity, and food intake, are displayed in Figure 3. The expected circadian profile is evident, establishing that subjects were indeed tested during all circadian phases. Therefore, the data in Figures 2 and 3 clearly confirm that the experimental design effectively induced fatigued performance due to the combined influence of sleep loss and circadian factors.

BOUT-TO-BOUT COHERENCE

Bout-to-bout coherence refers to the correlation between the total number of performance lapses in a 20-min. PVT bout and the results of a given drowsiness detection algorithm from a given technology. Since each volunteer subject studied had 20 bouts for comparison across the

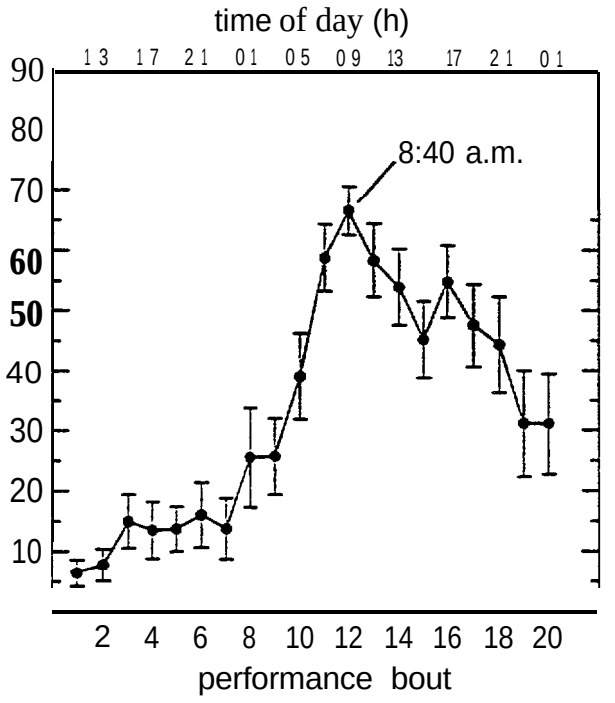

Figure 2. Mean (s.e.m.) number of psychomotor vigilance task performance lapses per 20-min. test bout across 20 test bouts (i.e., 42-hr of wakefulness) for the 14 subjects. PVT lapses increased significantly in the 42-hr period ($F_{19,247} = 23.72$, $p = 0.00001$).

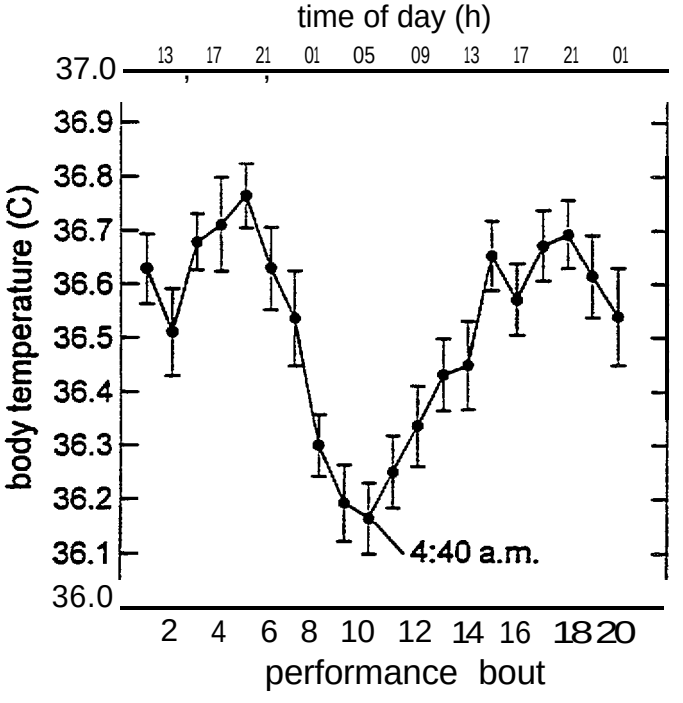

Figure 3. Mean (s-e-m.) oral temperature readings taken on 14 subjects at the end of each performance bout during 42-hr of wakefulness. A circadian profile is evident.

42-hr waking test period, bout-to-bout coherence was calculated within each subject for each technology that was available for the subject. In bout-to-bout coherence, the resulting coefficients reflect the extent to which a given technology/algorithm correlates with the total number of lapses in a 20-min. PVT bout. In other words, bout-to-bout coherence relies on a more global (i.e., 20-min.) estimate of vigilance across the 42-hr of waking, compared to minute-to-minute coherence (see below), which relies on vigilance detection at a higher temporal resolution.

Table 5 displays the bout-to-bout coherence coefficients (Pearson correlation coefficients) for individual subjects. Missing data in Table 5 and subsequent comparable tables are due to four major reasons: (1) Technologies were not available from suppliers for some subjects (CRI, ASC); (2) technologies interfered with acquisition of data from other technologies (MTI interfered with PERCLOS); (3) technologies had unreliable data storage/retrievability (IM); and (4) data remaining to be analyzed (SMM). Bout-to-bout coherence measures were available for all 14 subjects only for MTI's Alertness Monitor. The MTI column in Table 5 designates which subjects used different models of the Alertness Monitor. There were no statistically reliable differences in bout-to-bout coherence among any of the three MTI models (model 1 mean coherence r = 0.40 + 0.48; model 2 mean coherence r = 0.35 + 0.42; model 3 mean coherence r = 0.24 + 0.10; $F_{2,11}$ = 0.19, p = 0.82). Therefore for subsequent bout-to-bout analyses results from all three MTI models were pooled (n = 14).

Examination of Table 5 reveals a broad range of coherence coefficients within and between subjects and technologies (lowest coherence, r = - 0.54; highest coherence, r = 0.97). All but one of the technologies/algorithms had exceptionally high bout-to-bout coherence (i.e., r > 0.85) for at least one subject. However, for at least one subject, head position metrics (ASC60,

ASC90) and eye blink monitors (MTI, IM) also had either no measurable bout-to-bout coherence. or a significantly negative coherence, which resulted in a large inter-subject range of bout-to-bout coherence for these technologies/algorithms (i.e., inter-subject range > 0.64). Only the PERCLOS eye/facial ratings and the CRI EEG algorithm had relatively narrow inter-subject ranges (i.e.. <0.4) of positive bout-to-bout coherence among subjects.

Table 5. Bout-to-bout coherence for lapse frequency for individual subjects (Pearson correlation coefficients).

	eye/facial ratings			*EEG algorithms*		*heads position metrics*		*eye blink monitors*	
ID	**P70**	**P80**	**EM**	**CRI**	**SMM**	**ASC60**	**ASC90**	**MTI**	**IM**
6000	0.89	0.92	0.89	≠	.	0.83	0.82	0.10[1]	°
6001	0.85	0.83	0.84	≠	0.40	≠	≠	0.71[1]	0.20
6002	0.95	0.97	0.95	≠	.	0.91	0.85	0.90[1]	°
6004	0.84	0.83	0.83	≠	.	-0.54	0.20	0.54[2]	0.77
6005	0.94	0.94	0.95	0.54	.	≠	≠	-0.10[1]	°
6006	0.95	0.96	0.94	0.57	0.95	≠	≠	0.39[2]	0.54
6007	0.92	0.92	0.92	0.36	0.31	≠	≠	0.54[2]	0.32
6008	0.55	0.67	0.70	0.66	0.84	≠	≠	0.50[2]	0.85
6009	0.78	0.77	0.71	≠	.	0.23	0.13	0.67[2]	°
6011	0.95	0.97	0.95	≠	.	0.87	0.65	-0.48[2]	°
6014	*	*	*	≠	°	≠	≠	0.31[3]	0.79
6017	*	*	*	≠	°	≠	≠	0.34[3]	°
6019	*	*	*	≠	°	≠	≠	0.14[3]	°
6020	*	*	*	≠	°	≠	≠	0.17[3]	°

*PERCLOS data acquired but not storable due to glare from MTI glasses.
oData acquired but not retrievable due to hardware and/or software complications.
=Technology not available from supplier.
.Data remaining to be analyzed.
1MTI model 1 alertness monitor. 2MTI model 2 alertness monitor. 3MTI model 3 alertness monitor.

Graphic presentations of the subjects with the highest bout-to-bout coherence coefficient achieved (top graphs) for each technology and the lowest bout-to-bout coherence achieved (bottom graphs) for each technology, for the PVT lapse frequency criterion, are shown in Figure 4 (PERCLOS P80), Figure 5 (CRI EEG), Figure 6 (SMM EEG), Figure 7 (ASC90), Figure 8 (MTI), and Figure 9 (IM).

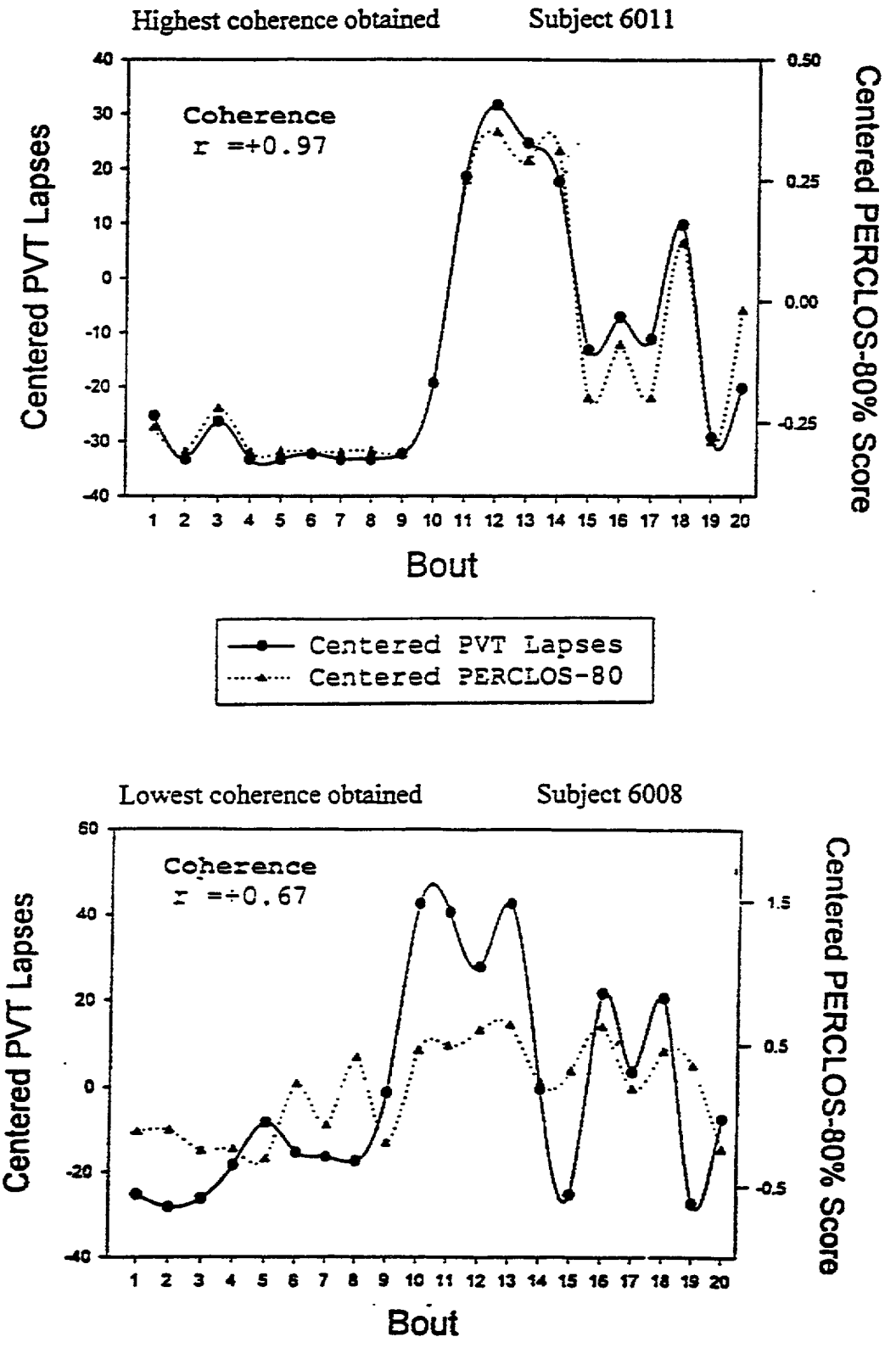

Figure 4. Coherence profiles for CMRI eye/facial rating ("PERCLOS 80"), for highest (top graph; subject 6011) and lowest (bottom graph; subject 6008) bout-to-bout coherence achieved for this technology/algorithm.

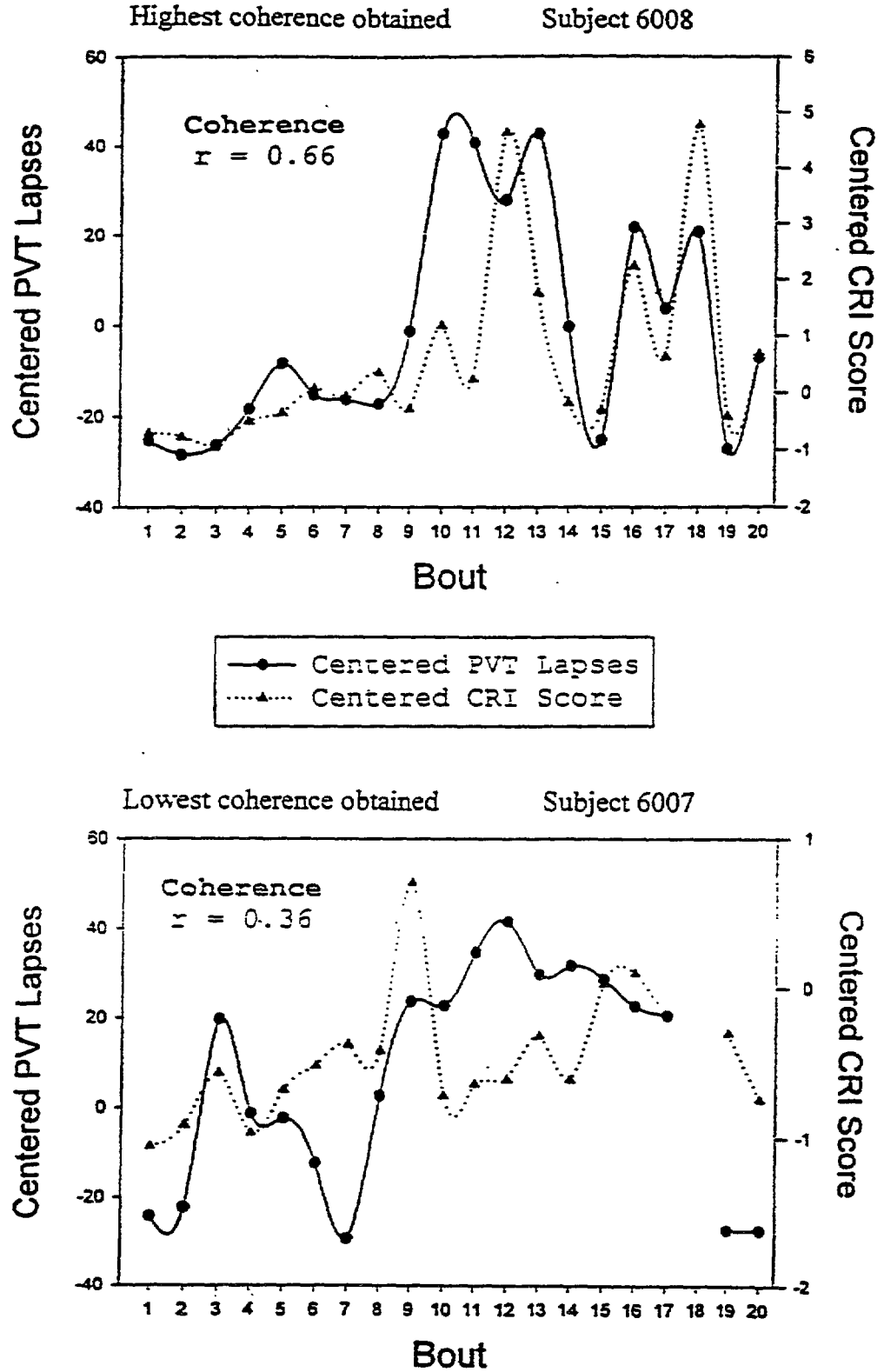

Figure 5. Coherence profiles for Consolidated Research, Inc. EEG algorithm ("Drowsiness Detection Algorithm"), for highest (top graph; subject 6008) and lowest (bottom graph; subject 6007) bout-to-bout coherence achieved for this technology/algorithm.

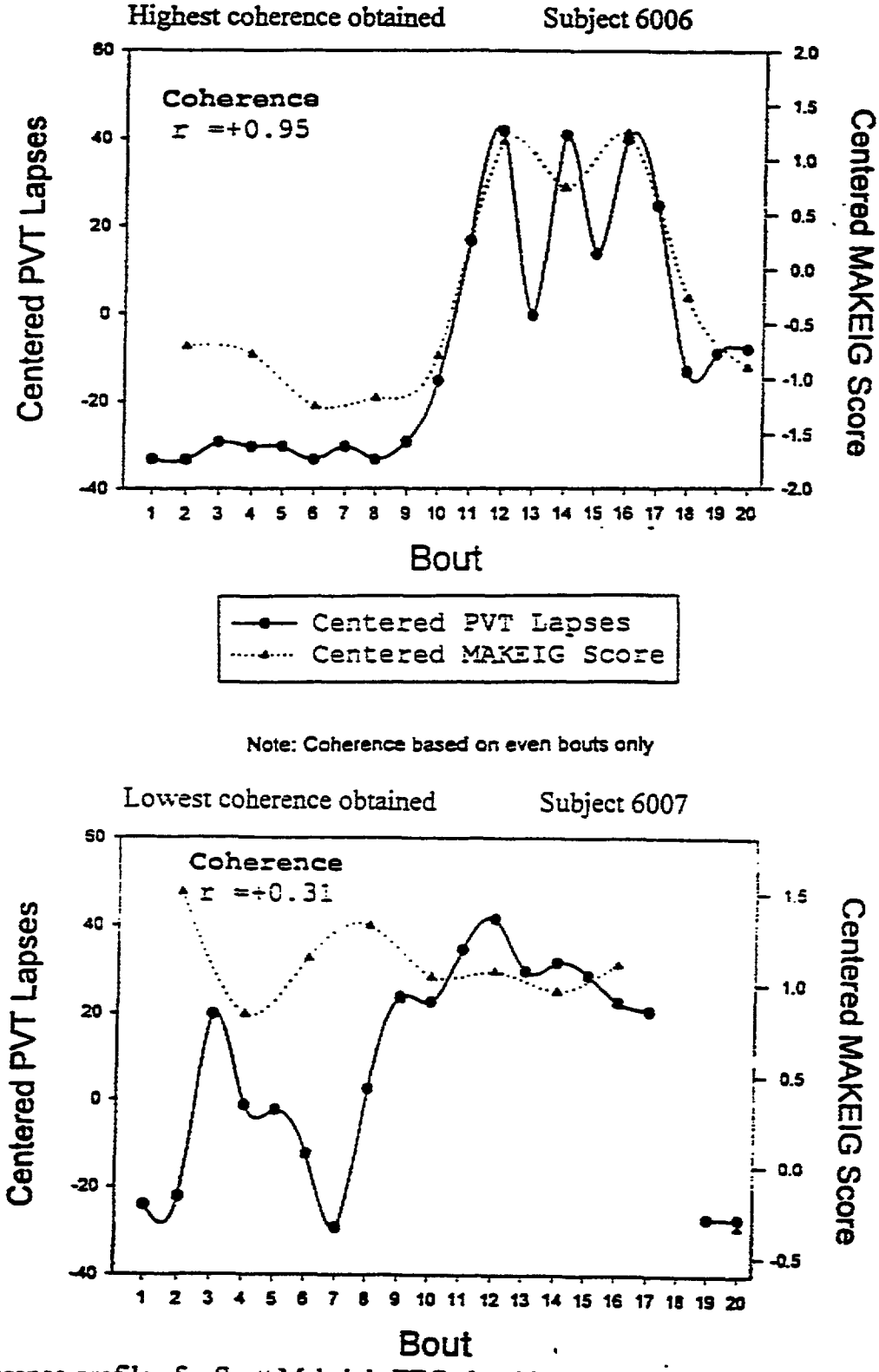

Figure 6. Coherence profiles for Scott Makeig's EEG algorithm, for highest (top graph; subject 6006) and lowest (bottom graph; subject 6007) bout-to-bout coherence achieved for this technology/algorithm.

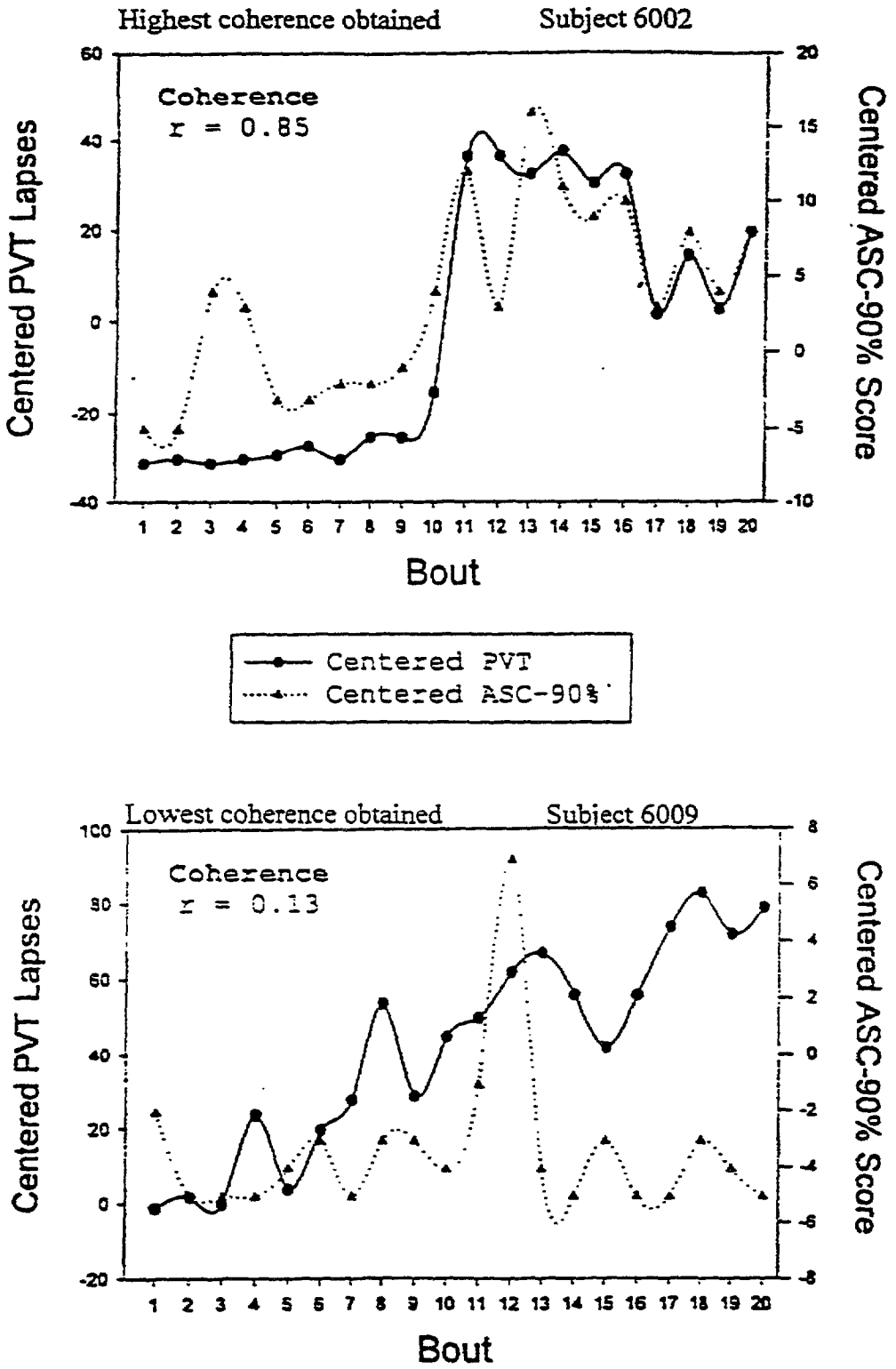

Figure 7. Coherence profiles for Advanced Safety Concepts, Inc. head position metric ("Proximity Array Sensing System"), for highest (top graph; subject 6002) and lowest (bottom graph; subject 6009) bout-to-bout coherence achieved for this technology/algorithm.

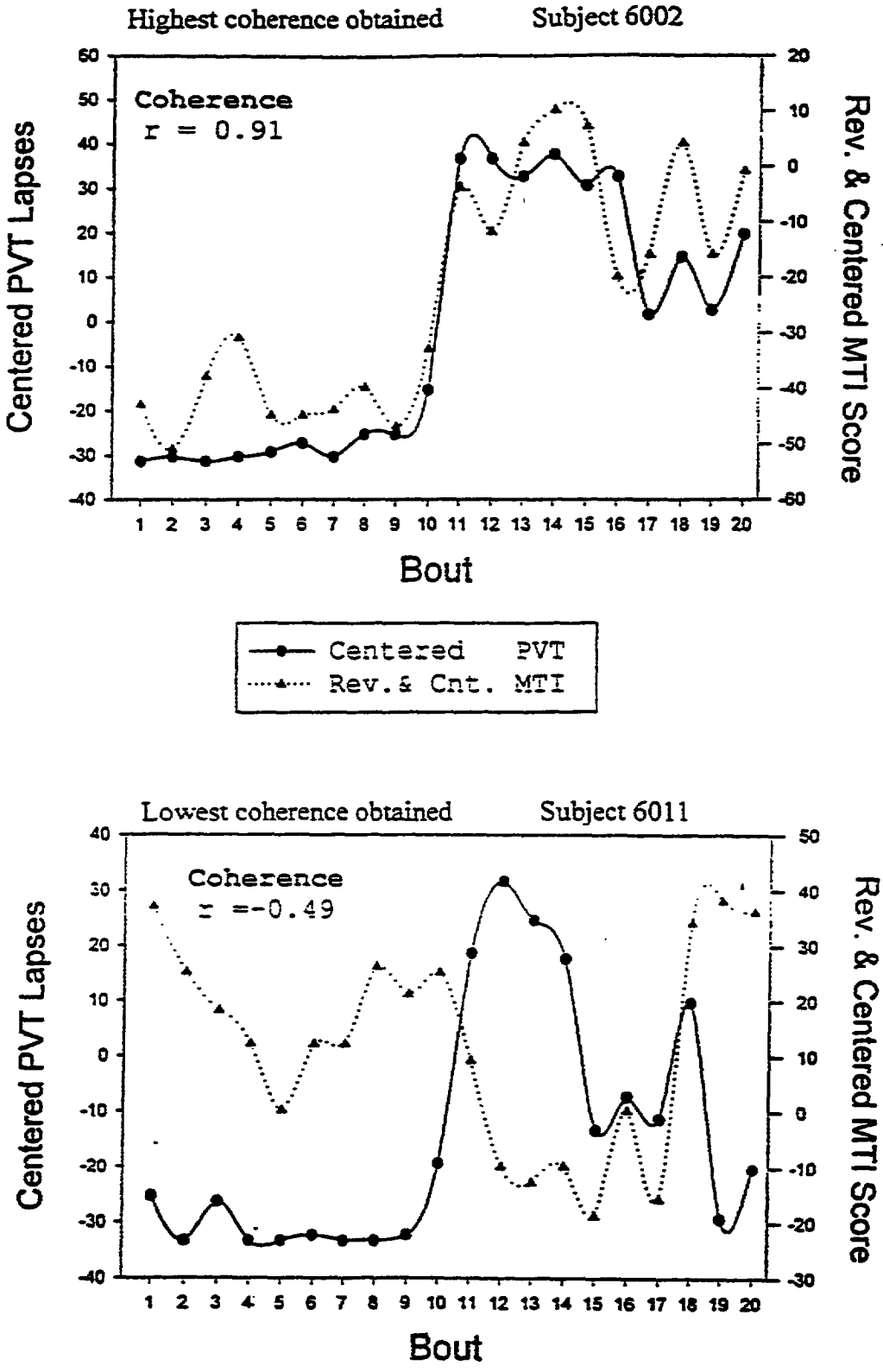

Figure 8. Coherence profiles for MTI Research, Inc. eye blink monitor ("Alertness Monitor"), for highest (top graph; subject 6002) and lowest (bottom graph; subject 6011) bout-to-bout coherence achieved for this technology/algorithm.

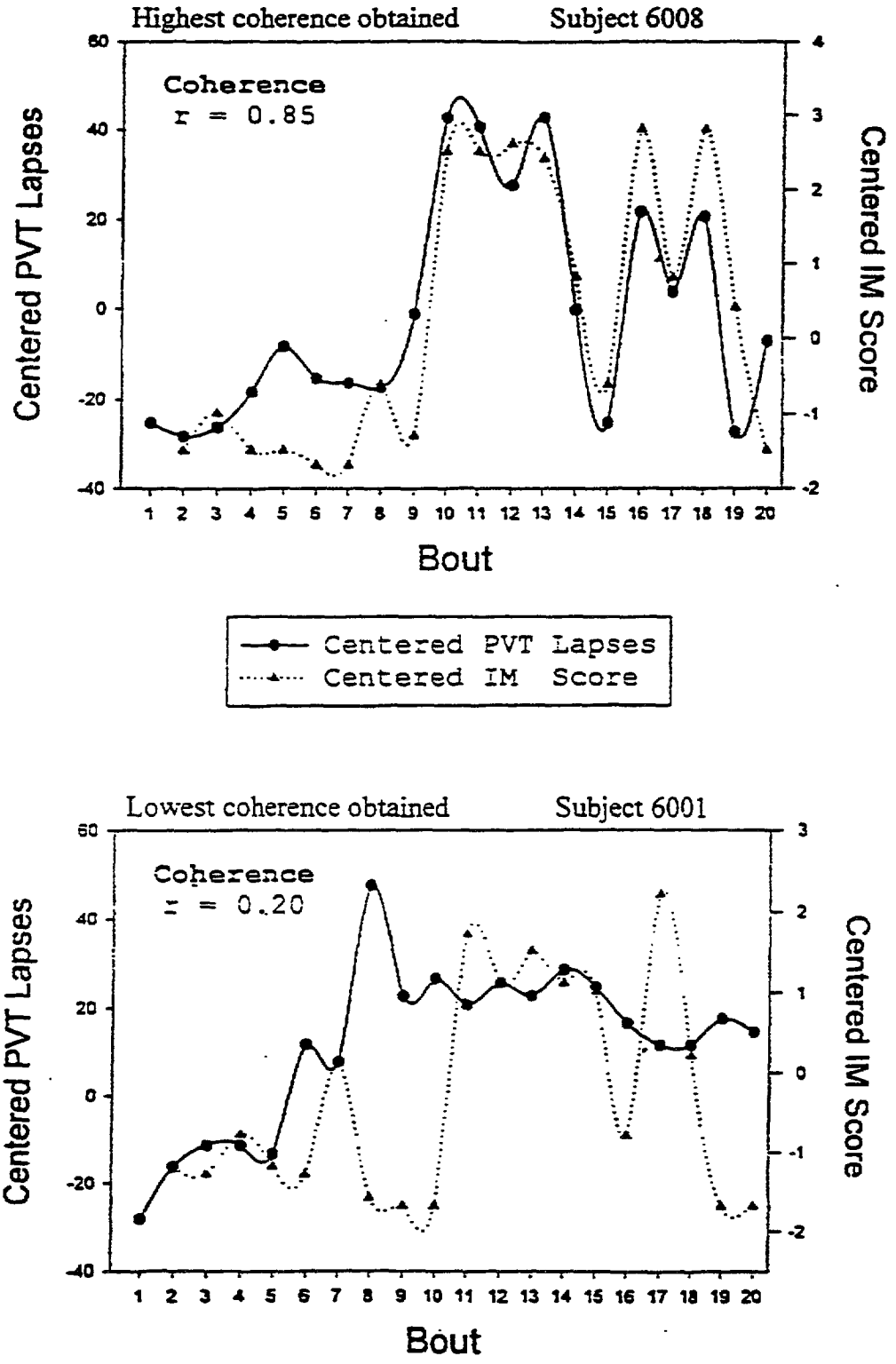

Figure 9. Coherence profiles for IM Systems, Inc. eye blink monitor ("Blinkometer"), for highest (top graph; subject 6008) and lowest (bottom graph; subject 6001) bout-to-bout coherence achieved for this technology/algorithm.

Table 6 displays the mean (SD) and median bout-to-bout coherence for lapse frequency for each technology. The average bout-to-bout coherence for eye/facial ratings PERCLOS variables were well above the average coherence values for all other technologies. PERCLOS metric P80 had the highest average bout-to-bout coherence (r = 0.875 + 0.10), which was also significantly greater than the bout-to-bout Pearson correlation between PVT lapses and subjects' visual analog ratings of their sleepiness taken immediately after each PVT trial (sleepiness VAS r = 0.626; t = - 3.9. p = 0.003). In other words, PERCLOS P80 correlated more highly with PVT lapses bout-to-bout than did subjects' own ratings of their sleepiness after performing the PVT.

Table 6. Average bout-to-bout coherence for lapse frequency (Pearson correlation coefficients).

	eye/facial ratings			*EEG algorithms*		*head position metrics*		*eye blink monitors*	
	P70	P80	EM	CRI	SMM	ASC60	ASC90	MTI	IM
N	10	10	10	4	4	5	5	14	6
Minimum	***0.55***	0.67	0.70	0.36	0.31	-0.54	0.13	-0.49	0.20
Maximum	***0.95***	***0.97***	***0.95***	0.66	0.95	0.91	0.85	0.91	0.85
Median	***0.90***	0.91	0.90	0.55	0.61	0.83	0.64	0.36	0.65
Mean	0.86	0.87	0.87	0.53	0.62	0.46	0.52	0.33	0.57
Std Dev.	***0.12***	0.10	0.09	0.12	0.31	0.62	0.34	0.36	0.27

No differences were found among the three highly intercorrelated PERCLOS measures (P70, P80, EM). Table 7 displays the average correlation among the bout-to-bout coherence measures for the various technologies. Only coherence for the head position metric ASC90 correlated significantly with coherence for each of the three PERCLOS metrics (r = 0.85 to 0.90, p = 0.063 to 0032), although the small sample sizes for most of the correlation coefficients shown in Table 7 severely limit the confidence that can be placed in these inter-technology relationships.

Table 7. Correlations among Pearson coefficients for bout-to-bout coherence for lapse frequency.

	MT1	CRI	IM	ASC60	ASC90	P70	P80	EM
SMM	-0.77	0.88	0.77			-0.30	-0.15	-0.21
	ns	ns	ns			ns	ns	ns
	4	3	4	0	0	4	4	4
EM	-0.38	-0.62	-0.62	0.63	0.86	0.90	0.96	
	ns	ns	ns	ns	0.058	0.0004	0.0001	
	10	4	5	5	5	10	10	
P80	-0.39	-0.60	-0.55	0.75	0.90	0.95		
	ns	ns	ns	ns	0.032	0.0001		
	10	4	5	5	5	10		
P70	-0.27	-0.64	-0.62	0.68	0.85			
	ns	ns	ns	ns	0.063			
	10	4	5	5	5			
ASC90	-0.24			0.84				
	ns		.	ns				
	5	0	1	5				
ASC60	-0.32		.					
	ns		.					
	5	0	1					
IM	-0.62	0.94						
	ns	ns						
	6	3						
CRI	-0.13							
	ns							
	4							

MINUTE-TO-MINUTE COHERENCE

Minute-to-minute coherence refers to the correlation between the total number of performance lapses in each minute of every 20-min. PVT bout across the 42-hr of waking, and the results of a given drowsiness detection algorithm from a given technology. Since each volunteer subject studied had a total of 20 PVT trials, and each bout was 20-min. in duration, 400

time points were available for calculating minute-to-minute coherence for each subject and technology. Table 8 displays the minute-to-minute coherence coefficients for lapse frequency (Pearson correlations) for individual subjects. Similar to bout-to-bout coherence data (see Table 5), there was a broad range of minute-to-minute coherence coefficients within and between subjects and technologies (lowest coherence, r = -0.33; highest coherence, r = 0.82). Minute-to-minute coherence measures were available for all 14 subjects only for MTI's Alertness Monitor. The MTI column in Table 8 designates which subjects used different models of the Alertness Monitor. There were no statistically reliable differences in minute-to-minute coherence among any of the three MTI models (model 1 mean coherence r = 0.29 + 0.29; model 2 mean coherence r = 0.23 + 0.32; model 3 mean coherence r = 0.15 + 0.15; $F_{2,11} = 0.23$, p = 0.79). Therefore for subsequent minute-to-minute analyses results from all three MTI models were pooled (n = 14).

Table 8. Minute-to-minute coherence for individuals subjects for lapse frequency (Pearson correlation coefficients).

	eye/facial ratings			*EEG algorithms*		*head position metrics*		*Eye blink monitors*	
ID	**P70**	**P80**	**EM**	**CRI**	**SMM**	**ASC60**	**ASC90**	**MTI**	**IM**
6000	0.68	0.73	0.68	≠	.	0.47	0.37	0.14[1]	°
6001	0.54	0.52	0.55	≠	0.27	≠	≠	0.37[1]	0.07
6002	0.78	0.77	0.82	≠	.	0.40	0.27	0.65[1]	°
6004	0.35	0.33	0.36	≠	.	-0.08	0.09	0.15[2]	0.24
6005	0.77	0.77	0.77	0.38	.	≠	≠	-0.02[1]	°
6006	0.79	0.80	0.81	0.38	0.73	≠	≠	0.16[2]	0.22
6007	0.74	0.73	0.75	0.12	0.36	≠	≠	0.57[2]	0.17
6008	0.34	0.40	0.43	0.27	0.49	≠	≠	0.38[2]	0.61
6009	0.39	0.38	0.37	≠	.	0.12	0.10	0.44[2]	°
6011	0.76	0.81	0.79	≠	.	0.59	0.48	-0.33[2]	°
6014	*	*	*	=	0	=	=	0.34[1]	0.46
6017	*	*	*	=	0	=	=	0.01[1]	0
6019	*	*	*	=	0	=	=	-0.01[3]	0
6020	*	*	*	=	0	=	=	0.11[3]	0

*PERCLOS data acquired but not scorable due to glare from MTI glasses.
"Data acquired but not retrievable due to hardware and/or software complications.
=Technology not available from supplier.
.Data remaining to be analyzed.
1MT1 model 1 alertness monitor; 2MTI model 2 alertness monitor; 3MTI model 3 alertness monitor.

Table 9 displays the mean (SD) and median minute-to-minute coherence for lapse frequency for each technology. As with bout-to-bout coherence, the average minute-to-minute coherence for eye/facial ratings PERCLOS variables were well above the average coherence values for all other technologies. As with bout-to-bout coherence, P80 had the highest average minute-to-minute coherence (r = 0.63 + 0.19).

Table 9. Average minute-to-minute coherence for lapse frequency (Pearson correlation coeffkients).

	eye/facial ratings			*EEG algorithms*		*head position metrics*		*eye blink monitors*	
	P70	**P80**	**EM**	**CRI**	**SMM**	**ASC60**	**ASC90**	**MTI**	**IM**
N	10	10	10	4	4	5	5	14	6
Minimum	0.34	0.33	0.36	0.12	0.27	-0.08	0.09	-0.33	0.07
Maximum	0.79	0.81	0.82	0.38	0.73	0.59	0.48	0.65	0.61
Median	0.71	0.73	0.72	0.32	0.43	0.40	0.27	0.17	0.23
Mean	0.61	0.63	0.63	0.29	0.46	0.30	0.26	0.22	0.29
Std Dev	0.18	0.19	0.18	0.12	0.20	0.27	0.16	0.26	0.20

Although PERCLOS had both higher bout-to-bout and minute-to-minute coherence than all other technologies/algorithms, there was a fundamental difference within each technology/algorithm between minute-to-minute and bout-to-bout coherence. Minute-to-minute coherence for lapse frequency was consistently lower than bout-to-bout coherence for lapse frequency. These comparisons are displayed in Table 10. The differences between bout-to-bout coherence and minute-to-minute coherence reached statistical significance for all three PERCLOS measures (p = 0.000l), the CRI EEG measure (p = 0.0l), and both the MTI (p = 0.02) and IM (p = 0.005) eye blink monitors. There was a trend for a reliable difference for the ASC90 head position metric (p = 0.06), and the differences found for the SMM EEG measure and ASC60 head position metric were in the same direction. It appears that nearly all technologies

had a better prediction of PVT lapses when sampling involved a longer (i.e., 20 min.) rather than a briefer (i.e., 1 min.) time period.

Table 10. Comparison of bout-to-bout and minute-to-minute coherence measures for lapse frequency.

	eye/facial ratings			*EEG algorithms*		*head position metrics*		*eye blink monitors*	
	P70	**P80**	**EM**	**CRI**	**SMM**	**ASC60**	**ASC90**	**MTL**	**IM**
Number Ss studied	10	10	10	4	4	5	5	14	6
By Bout 1 - 20	0.86	0.87	0.87	0.53	0.62	0.46	0.52	0.33	0.57
By Minute 1 - 20	0.61	0.63	0.63	0.29	0.46	0.30	0.26	0.22	0.29
Mean Difference	0.25	0.24	0.24	0.24	0.16	0.16	0.26	0.11	0.28
t =	**-6.9**	-6.7	-6.8	-4.5	-1.9	-0.9	-2.4	-2.4	-4.7
p =	0.0001	0.0001	0.0001	0.01	ns	ns	0.06	0.02	0.005

PERCLOS COHERENCE AS A FUNCTION OF TIME BASE

As noted above, the differences between bout-to-bout coherence and minute-to-minute coherence for lapse frequency were statistically significant for all PERCLOS measures. This raises the question of the coherence that could be achieved with PERCLOS as a function of varying durations of time between 1-min. and 20-mins. In an effort to determine whether intervals less than 20 min. but greater than 1 min. could yield coherence comparable to that found for 20 min., coherence was calculated for the PERCLOS P80 metric for 6 separate time,. base intervals (1 min., 2 min. 4 min., 5 min., 10 min., and 20 min.). Intervals of 2,4,5, and 10 minutes were selected because of previous data suggesting that the greatest increments in coherence could be expected between 1 and 6 min. (W. Wierwille, personal communication, January 6, 1998), and because these values are multiples of 20 min., which prevents loss of data in the calculations.

Not surprisingly, an analysis of variance on the resulting coherence values across the six time-base calculations of PERCLOS was statistically significant ($F_{4,45} = 34.67$, $p = 0.00001$). Least squares polynomial analysis revealed both linear ($F_{1,9} = 37.63$, $p = 0.00001$) and cubic ($F_{1,9} = 42.95$, $p = 0.00001$) trends. Figure 10 displays these results graphically for the average coherence obtained for PERCLOS P80 drowsiness metric as a function of the sampling time base. A distance-weighted least squares function was fit to the data in Figure 10. Paired t-test comparisons between adjacent pairs of points in Figure 10 revealed statistically significant increments in coherence with each increment in the time base (min. 1 vs. 2, $t = -14.28$, $p = 0.0001$; min. 2 vs. 4, $t = -7.73$, $p = 0.0001$; min. 4 vs. 5, $t = -3.08$, $p = 0.013$; min. 5 vs. 10, $t = -3.42$, $p = 0.008$; min. 10 vs. 20, $t = -4.79$, $p = 0.001$). It appears that optimization of PERCLOS coherence for lapse frequency requires a time above 1 minute (c.f., Appendix).

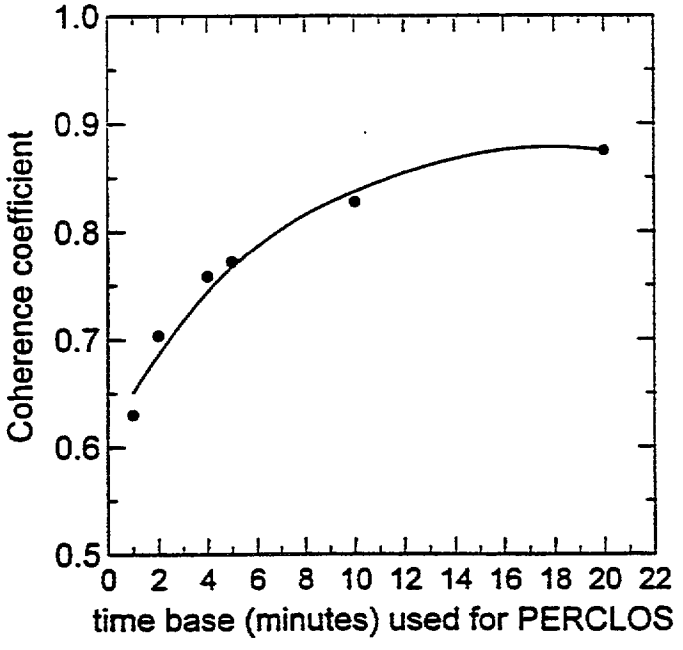

Figure 10. Mean PERCLOS P80 coherence across 42-hr of waking, as a function of the time base used to define an epoch. A distance-weighted least squares function was fit to the data.

COHERENCE VARIABILITY

Intra-Subject Variability in Coherence: Day 1 vs. Day 2 of Waking

To determine whether each bout-to-bout coherence was comparable across the broad range of sleepiness/alertness induced by the within-subjects experimental design, coherence was also calculated separately for the first 22-hr of wakefulness (i.e., performance bouts 1 to 10; from 10:00 a.m. on day 1 to 4:00 a.m. on day 2), and compared to coherence for the final 20-hr of waking (i.e., performance bouts 11 to 20; from 6:00 a.m. on day 2 to just after midnight on the start of day 3), when subjects were much sleepier and lapsed more frequently (see Figure 2). Table 11 displays the results of these analyses for both bout-to-bout and minute-to-minute coherences for lapse frequency. There were no statistically reliable differences in coherence for bouts 1 - 10 versus bouts 11 - 20. There was a trend for PERCLOS bout-to-bout coherence to be higher during the first 22-hr of waking (e.g., EM average r = 0.85) than during the final 20-hr of waking (EM mean r = 0.63, t = - 1.8, p = 0.10). This suggests that PERCLOS was predictive of PVT lapses in a range of wakefulness commonly experienced by motor vehicle operators (i.e., < 22 hr awake).

Table 11. Coherence measures for lapse frequency for bouts #l to10 vs. bouts #1 1 to 20 (Pearson coefficients).

	Eye/facial ratings			*EEG algorithms*		*head position metrics*		*eye blink monitors*	
	P70	**P80**	**EM**	**CRI**	**SMM**	**ASC60**	**ASC90**	**MTL**	**IM**
Number Ss Studied =	10	10	10	4	4	5	5	14	6
By Bout									
1-10 (2-22 hr awake)	*0.83*	*0.85*	*0.80*	*0.22*	0.30	*0.35*	*0.33*	*0.19*	*0.15*
11-20 (22-42 hr awake)	**0.64**	**0.65**	**0.63**	**0.33**	**0.72**	**0.54**	**0.50**	**0.10**	**0.36**
Paired t-test t =	*-1.8*	*-1.6*	**-1.8**	**0.4**	**1.0**	**0.6**	**0.6**	**-0.6**	**0.8**
p =	**0.11**	**0.13**	**0.10**	*Ns*	**ns**	**ns**	**ns**	ns	ns
By Minute									
1-10 (2-22 hr awake)	0.40	0.42	0.43	0.03	0.32	0.17	0.14	0.14	0.15
11-20 (22-42 hr awake)	0.48	0.49	0.50	0.23	0.44	0.23	0.21	0.17	0.15
Paired t-test t =	0.9	0.8	0.9	1.8	0.6	0.7	0.8	0.4	0.0
p =	ns	Ns	ns	Ns	ns	ns	ns	ns	ns

Inter-Subject Variability in Coherence: Lower Lapsers vs. Higher Lapsers

As expected, all subjects eventually became sleepy and evidenced increased PVT lapsing during the course of the 42-hr period of waking. However, as is commonly observed in many laboratory sleep-deprivation experiments (Dinges & Kribbs, 1991), a subset of the subjects accounted for a disproportionate number of the lapses. This observation has also been reported in results from field studies of 24-br transportation operations (Rosekind et al, 1994), and it was found in the recent North American study of fatigue in 80 over-the-road commercial motor vehicle operators, 14% of whom accounted for 54% of the observed drowsy driving episodes (Wylie et al., 1996). Therefore, a fundamental issue in alertness detection in drivers is the extent to which technologies/algorithms can detect both average levels of drowsiness, and the subset of persons who have more severe drowsiness responses.

To determine whether bout-to-bout coherence was systematically affected by the tendency of a subject to lapse during sustained waking, an analysis of the distribution of PVT lapses by subjects was performed to segregate two groups for comparison based on their tendency to lapse: (1) those subjects with less than average tendencies to lapse as determined by PVT lapse $z < 0$ (see Statistical Methods section); and (2) those subjects with greater than average tendencies to lapse (PVT lapse $z > 0$). Two-sample t-tests were used to compare technology/algorithm coherence values between "higher lapsers" (HL; n = 6 subjects who averaged > 38 PVT lapses per bout), and "lower lapsers" (LL; n = 8 subjects who had an average of < 33 PVT lapses per bout). HL subjects were 42% of the subjects, but they accounted for 976 lapses, or 69% of all PVT lapses recorded in the study. Figure 11 displays the mean (SD) total number of PVT lapses for LL subjects compared to HL subjects during the first 22-hr of waking and the final 20-hr of waking. Figure 11 makes clear that the differences between the two groups

are present during both the first and second halves of the study. Remarkably, the mean PVT lapse total for the HL subgroup during the first 22 hr of waking is at a level comparable to that of the LL subgroup after a night without sleep. These differences were not associated with differential sleep histories. Analyses of the actigraphic and sleep diary records revealed no statistically significant differences in either sleep quality or average nocturnal sleep duration during the 4 nights prior to laboratory (LL = 7.13 ± 0.33 hr; HL = 7.37 ± 0.42 hr).

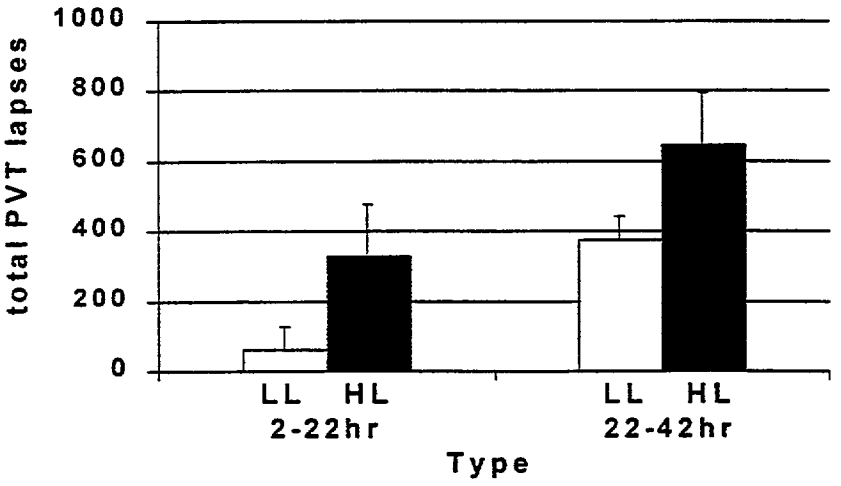

Figure 11. Mean (SD) of total number of PVT lapses for 8 lower lapser [LL] subjects compared to 6 higher lapser [HL] subjects during the first 22-hr of waking and the final 20-hr of waking.

Table 12 compares bout-to-bout coherence results for lapse frequency for LL and HL subgroups for each technology/algorithm by individual subjects and subgroup means. The "ZPVT" parameter in the Table represents the segregation analysis criterion. Statistical comparisons between LL and HL subgroups were carried out for the average bout-to-bout coherence for lapse frequency for PERCLOS P80 (P70 and EM measures were highly

intercorrelated with P80), for ASC90, and for MTI. (Other technologies/algorithms had too few subjects in one or both of the subgroups to warrant analyses.)

Table 12. Bout to bout coherence for lapse frequency for lower lapsers and higher lapsers.

ID	*lapse total per bout*	*ZPVT*	*total sleep time 4 days prior to TSD*	*eye/facial ratings*	*EEG algorithms*		*head position metrics*	*eye blink monitors*	
Lower lapsers				**P80**	**CRI**	**SMM**	**ASC90**	**MTI**	**IM**
6000	16.70	-0.95	7.00	0.92	≠	.	0.82	0.10	°
6002	31.80	-0.09	7.67	0.97	≠	.	0.85	0.91	°
6005	18.35	-0.85	6.92	0.94	0.54	.	≠	-0.10	°
6006	25.70	-0.44	7.58	0.96	0.57	0.95	≠	0.39	0.54
6008	32.45	-0.05	7.08	0.67	0.66	0.84	≠	0.50	0.85
6011	19.25	-0.80	7.00	0.97	≠	.	0.65	-0.49	°
6017	26.80	-0.37	7.17	*	≠	°	≠	0.34	°
6019	3.90	-1.68	6.67	*	≠	°	≠	0.14	°
Mean =	21.86	-0.65	7.13	0.90	0.58	0.89	0.77	0.22	0.69
SD =	9.39	0.53	0.33	0.11	0.06	0.08	0.10	0.41	0.22
.									
Higher lapsers									
6001	44.85	0.65	7.17	0.83	≠	0.40	≠	0.71	0.20
6004	50.50	0.95	7.33	0.83	≠	.	0.20	0.54	0.77
6007	40.40	0.39	7.25	0.92	0.36	0.31	≠	0.54	0.32
6009	75.30	2.39	6.83	0.77	≠	.	0.13	0.67	°
6014	43.20	0.56	7.58	*	≠	°	≠	0.31	0.79
6020	38.70	0.30	8.08	*	≠	°	≠	0.17	°
Mean =	48.82	0.87	7.37	0.83	0.36	0.35	0.16	0.49	0.51
SD=	13.59	0.77	0.42	0.06	--	0.06	0.04	0.21	0.30

*PERCLOS data acquired but not scorable due to glare from MTI glasses.
"Data acquired but not retrievable due to hardware and/or software complications.
*Technology not available from supplier.
.Data remaining to be analyzed.

The MTI eye-blink monitor tended to have a lower bout-to-bout coherence for the 8 LL subjects (mean coherence r = 0.22) compared to the 6 HL subjects (mean coherence r = 0.49; t = - 1.57, df = 12, p = 0.145). In contrast, head position metric ASC90 only had a couple of subjects in each subgroup, but it yielded a significantly higher bout-to-bout coherence for the 3 LL subjects (mean coherence r = 0.77) compared to 2 HL subjects (mean coherence r = 0.16; t =

8.63, df = 3, p = 0.0057). Importantly, PERCLOS P80 yielded comparably high bout-to-bout coherence between the 6 LL subjects (mean coherence r = 0.90) and the four HL subjects (mean coherence r = 0.83; t = 1.07, df = 8, p = 0.316) on which PERCLOS data were available. Thus, despite large inter-subject variability in PVT lapse rates, PERCLOS P80 appeared to reliably predict bout-to-bout lapsing across LL and HL subgroups.

PERCLOS: PREDICTIVE VALUE, SENSITIVITY, SPECIFICITY

As described in the sections above, the three eye/facial ratings of PERCLOS were highly intercorrelated, but PERCLOS measures consistently averaged higher bout-to-bout coherence for lapse frequency ($r = 0.875 \pm 0.10$) than all other technology/algorithm measures, and than sleepiness ratings made by subjects during the performance bouts. Consequently, positive and negative predictive values, and sensitivity and specificity were calculated for PERCLOS variable P80. In this set of exploratory analyses, the criteria for defining a bout as "truly drowsy" were based on whether total PVT lapses were larger or smaller than the grand mean computed over all subjects and bouts (grand mean = 33). To determine the sensitivity of results for this definition, computations were repeated defining a bout as truly drowsy if its total PVT lapses was greater than 1.2 times this grand mean value, and then again if its value was 0.80 times the grand mean. Similarly, the criterion for predicting that a subject was drowsy during a particular bout was based on whether the PERCLOS P80 score for that bout was larger or smaller than the grand mean over all bouts and subjects (grand mean = 0.31). Again, to determine the sensitivity of our results on the prediction criterion, we varied the prediction rule using 1.2 times and 0.80 times the P80 grand mean. Table 13 displays the mean (SD) positive and negative predictive values and sensitivity and specificity of P80 as a function of different combinations of the PVT lapse and P80 drowsiness criteria for the n = 10 subjects with P80 data.

Table 13. Mean (SD) positive and negative predictive values, sensitivity and specificity of P80.

PVT criterion: P80 criterion*	Positive predictive value %	Negative predictive Value %	Sensitivity %	Specificity %
1.0 : 1.0	87.9 (12.9)	88.7 (9.8)	88.6 (8.7)	90.6 (10.5)
1.0 : 0.8	88.3 (12.6)	92.7 (8.9)	92.6 (9.0)	90.6 (10.5)
1.0 : 1.2	88.4 (13.5)	79.1 (12.7)	75.6 (9.6)	92.6 (8.9)
1.2 : 0.8	73.0 (16.9)	98.9 (3.5)	99.0 (3.2)	77.6 (16.0)
1.2 : 1.2	80.6 (12.8)	94.4 (6.7)	90.7 (10.1)	87.9 (7.9)
0.8 : 0.8	93.8 (11.2)	85.4 (10.8)	87.8 (7.8)	94.4 (12.5)
0.8 : 1.2	93.6 (11.4)	72.8 (16.2)	71.9 (12.2)	95.4 (9.5)

* The criteria listed are relative to the grand means. That is, 1:1 means that a bout was defined as a "truly drowsy" bout if the total number of PVT lapses was larger than the grand mean over all bouts (PVT lapses >33). Similarly, a bout was predicted to be a "drowsy bout" if its P80 value was larger than the grand mean of all P80 bout values (P80>0.31). These grand means were computed over all subjects and bouts. The values of 0.8 and 1.2 refer to criterion values that are .80 as large or 1.2 larger that these grand means.

Positive predictive value is defined as the percentage of bouts predicted to have come from subjects when they are drowsy that are, in fact, truly drowsy bouts. Thus, the base-case value in Table 13 of 87.9% (PVT:P80 at 1:1) indicates that on average, of all bouts predicted to be drowsy (i.e., predicted to have total PVT lapses greater than the grand mean), 87.9% of these bouts actually did have PVT lapses > 33. Negative predictive values is defined as the percentage of bouts predicted to have come from subjects when they are not drowsy that are, in fact, truly not drowsy bouts. Thus, the base-case value in Table 13 of 88.9% (PVT:P80 at 1:1) indicates that, on average, of all bouts predicted not to be drowsy (i.e., predicted to have total PVT lapses less than the grand mean), 88.9% of these bouts actually did have PVT lapses less than the grand mean.

The prediction analysis summarized in Table 13 revealed that depending upon the criteria used to define whether a bout truly came from a subject when he was drowsy and depending upon the specific prediction rule, positive predictive value pairs ranged from 73.0% to 93.6%, using (PVT, P80) criteria of (1.2:0.80) and (0.80,1.20), respectively, while negative predictive

value pairs ranged from 98.9% to 72.8%. The upper bound values suggest that further research fine-tuning the definitions of true and predicted drowsy states based on varying P80 cut-points and the definition of drowsiness (PVT lapses) as well as appropriately taking into account costs associated with false positives and false negatives may result in economically relevant strategies for signaling elevated risk for performance decline, while minimizing false alarms.

Table 13 also displays sensitivity and specificity values for the various PVT:P80 criterion ratios. In general, these values are not useful for summarizing a prediction rule's ability to predict correctly. Sensitivity is defined as the proportion of truly drowsy bouts that are predicted to be drowsy, while specificity is defined as the proportion of truly non-drowsy bouts that are predicted not to be drowsy. They are provided because positive and negative predictive values depend greatly on prevalence while, in theory, sensitivity and specificity do not. The proportion of bouts defined as drowsy (roughly 50% of bouts had total PVT lapses >33) depends upon our experimental paradigm and the population from which the sample was drawn. The sensitivity and specificity values provided in Table 13 may be useful when used in conjunction with Bayes Rule to determine how the positive and negative predictive values would vary assuming drowsy bout prevalences that are larger or smaller than the 50% value in our sample.

EXPERIMENT I: DISCUSSION AND CONCLUSIONS

The 42-hr sleep-deprivation experimental paradigm coupled with quasi-continuous psychomotor vigilance performance testing was effective in producing a wide range of variation in vigilance lapse rates within and between subjects, consistent with published literature (Kribbs & Dinges, 1994; Dinges et al., 1994; Dinges et al., 1997). The nine drowsiness metrics we tested from a total of six technologies also showed substantial variation within and between subjects, and in their coherence with PVT lapses. Only one of these technologies, eye/facial ratings of PERCLOS (and its three drowsiness metrics), consistently covaried with PVT lapses across the 42-hr (Tables 5 and 6). Unlike other technologies, PERCLOS relied on human observers' ratings of subject's faces during PVT performance, and in particular, on observers' ratings of subjects' slow eyelid closures (rather than blinks), according to a system developed by Wierwille and colleagues (Wierwille et al., 1994; Wierwille & Ellsworth, 1994). Coherence coefficients for the three PERCLOS drowsiness metrics (i.e., P70, P80, EYEMEAS) were highly intercorrelated and had only a trivial average difference in coherence (mean = 0.02) among each other across 10 subjects. Although this suggests that the three PERCLOS metrics are redundant, it does not diminish the remarkable accuracy PERCLOS had in reflecting PVT lapses. PERCLOS had an average bout-to-bout coherence for $n = 10$ subjects of $r = 0.875 \pm 0.10$ (Table 6), with 60% of subjects ≥ 0.92 (Table 5). No other technology came close to this level of coherence. Moreover, the average high coherence obtained in the current study ($r = 0.875$) is nearly identical to the average validation coefficient obtained by Wierwille and colleagues (1994) in the original validation study of PERCLOS on $n = 12$ subjects ($r = 0.872$). This agreement is all the more remarkable given that Wierwille et al. (1994) used simulated driving variables and an auditory

vigilance task to establish PERCLOS validity, while the current study used psychomotor vigilance lapses as the validation criterion. In short, PERCLOS appears to predict hypovigilance robustly on both auditory and visual vigilance tasks as well as simulated driving.

In the current study, PERCLOS not only outperformed other drowsiness-detection technologies, but it also correlated more highly with PVT lapse frequency bout-to-bout than did subjects' own ratings of their sleepiness after performing the PVT ($t = -3.9$, $p = 0.003$), and it had high positive and negative predictive values (Table 13). There was evidence that PERCLOS was predictive of lapses in the first 22-hr of waking (Table 11; mean $r = 0.872$)--a time frame that should apply to the majority of drivers. The power of PERCLOS to predict PVT lapses was also robust relative to marked individual differences in lapse tendencies (Table 12). Predicting hypovigilance in the face of substantial individual differences is an important and highly promising outcome, since substantial inter-subject variability in drowsiness was observed in the recent USA-Canada driver fatigue and alertness study, where 14% of drivers accounted for 54% of all observed video-drowsiness episodes (Wylie et al., 1996).

The present findings in conjunction with those of Wierwille et al. (1994; Wierwille & Ellsworth, 1994) suggest that PERCLOS has the potential to detect fatigue-induced lapses of attention (i.e., hypovigilance) and drowsiness-related interruptions of visual input (i.e., diminished visual awareness) during driving, if the following can be achieved. (1) The PERCLOS scoring algorithm used by human observers in laboratory studies must be automated in a computer algorithm with demonstrated evidence of acceptable levels of validity and reliability. (2) PERCLOS must be validly and reliably measured during driving, using unobtrusive technologies (e.g., video image analysis, infrared eye tracking). (3) Acceptable levels of positive and negative predictive values for driver fatigue must be determined for an

automated, over-the-road version of PERCLOS. Attempts to meet the above criteria in order to transition an on-line, automated version of PERCLOS to a realistic over-the-road environment are currently underway at Carnegie Mellon Research Institute (Grace et al., in preparation), at LC Technologies (reported by W. Wierwille, January 23, 1998), and at Applied Science Group, Inc. (reported by W. Rogers, February 11, 1998) . There are also video-based drowsy-driving detection systems under various stages of development at motor vehicle manufacturers (Fukuda et al., 1995; Kaneda et al., 1994; Richardson, 1995).

In addition to high bout-to-bout coherence and predictive values, PERCLOS yielded the highest average coherence of all six technologies when the temporal window used to calculate coherence across the 42-hr period of waking was reduced from 20-minute epochs to 1 -minute epochs (Tables 8 and 9). However, PERCLOS minute-to-minute (1 minute epoch) coherence was statistically significantly below its bout-to-bout (20-minute epoch) coherence (e.g., P80 mean minute-to-minute r = 0.63 +0.19 vs. mean bout-to-bout r = 0.87 +0.10; t = - 6.7, p = 0.0001). This decrease in coherence at l-minute epoch lengths was also evident for the other technologies and drowsiness metrics (Table 10). It suggests that there is error in biobehavioral measurement of hypovigilance in any given minute, but that summing over a 20-minute period reduces this error and a more accurate measure of drowsiness is obtained. Not surprisingly, the performance-impairing effects of drowsiness become more evident as the time-base (sampling base) for their assessment increases. This was clearly the case for PERCLOS (Figure 10). Although it is well established scientifically that the level of sleepiness changes dramatically over hours as a function of the interplay of two biological processes (wake duration and endogenous circadian phase), the factors that systematically relate to minute-to-minute waxing and waning of

hypovigilance are not understood. Until these hypothesized high frequency modulators of drowsiness/alertness are identified and made measurable, it appears that technology to track hypovigilance in drowsy drivers will need to use a temporal epoch longer than 1 minute, and probably as long as 20 minutes.

Although on average PERCLOS predicted psychomotor vigilance performance lapses better than all other technologies, most of the five technologies had bout-to-bout coherence for lapse frequency values for at least one subject comparable to what was found for all 10 subjects using PERCLOS (Figures 5,6,7,8,9). While the 5 non-PERCLOS technologies predicted lapses well on at least one subject each, it is interesting that a subject for which a high bout-to-bout coherence was found on one technology did not necessarily have a high bout-to-bout coherence on another technology, even when the two technologies were in the same domain (e.g., both EEG algorithms, or both eye blink monitors). Thus, the failure to find high coherence on a given subject for a given technology was not due to the subject having immeasurable drowsiness. Even subject 6008, for which the three PERCLOS metrics yielded their lowest bout-to-bout coherence ($r = 0.55$ to $r = 0.70$), was more reliably predicted by Dr. Scott Makeig's EEG algorithm ($r = 0.84$) and IM System's Blinkometer ($r = 0.85$). This suggests that a combination of two highly valid technologies may further enhance the detection of hypovigilance, although the utility of this approach would depend on demonstrating that one technology could consistently track drowsiness when they other was less accurate, and vice versa.

While each technology showed relatively high bout-to-bout coherence for lapse frequency for at least one subject, with the exception of PERCLOS, each technology also yielded low coherence on at least one subject and the specific subject for which this occurred depended on the technology (Table 5). The two EEG algorithms, CRI and SMM, performed the best of the

non-PERCLOS technologies in this regard. They had lowest coherence coefficients of r = 0.36 and r = 0.3 1, respectively. In comparison, one of the head position metrics (ASC60) and MTI's eye blink device produced substantial negative coherence coefficients on different individual subjects (r = - 0.54 and r = - 0.49, respectively), suggesting that to some extent for these two specific subjects their indices of drowsiness varied inversely with PVT lapses (i.e., more drowsiness = fewer PVT lapses).

It is possible that specific characteristics of the validation task (i.e., PVT performance lapses) may have favored PERCLOS and/or altered the biobehavioral markers for other technologies. Thus, reaction times to signals would be expected to be longer the greater proportion of time the eyelids were closed, so finding a high coherence between PVT lapses and PERCLOS is not surprising in retrospect. In addition, the specific task demands of the PVT may have suppressed eye blinks, or altered the EEG or head position metrics relative to what night be observed while driving. There is no evidence that the sustained attentional demands of the PVT task are fundamentally different from the sustained attentional demands of driving. Moreover, systematic biasing of outcomes by PVT task demands cannot explain why all of the technologies could achieve high coherence on at least one subject, and for most of them this occurred on 2-3 subjects (Table 5). To the extent that each technology was capable of predicting PVT lapses from at least one subject, all of the technologies tested displayed some degree of potential as hypovigilance detectors. It therefore it remains possible for some of these technologies to further improve their "detection" of lapses. With this in mind, PVT lapse data have been sent to each supplier (after the results of this prospective study were established), to permit a retrospective 'tweaking" of drowsiness algorithms for the purpose of enhancing their detection of lapses. However, a cautionary note is in order. If such a retrospective "enhancement" of coherence

proves possible for some of the technologies, a prospective re-validation test 'of the 'tweaked" drowsiness algorithm would be necessary.

It is also important to keep in mind that data were available for only 4 subjects for each of the two EEG algorithms (CRI and SMM), and for only 5 subjects for the two drowsiness metrics from the head position monitor (ASC60, ASC90). Therefore it is open to question whether the similar moderate average coherences for lapse frequency of these technologies are reliable (CRI r = 0.53 + 0.12; SMM r = 0.62 + 0.31; ASC60 r = 0.46 + 0.62; ASC90 r = 0.52 + 0.34). In contrast, data were acquired on all 14 subjects for the two eye blink monitors (MT1 and IM), and on 10 subjects for PERCLOS. The result for PERCLOS was uniformly high coherence, while the result for MTI's device was at the other end of the spectrum, averaging the lowest bout-to-bout coherence for lapse frequency (r = 0.33 + 0.36; Table 6). There were no statistically reliable differences in bout-to-bout coherence for lapse frequency among any of the three MTI models (1st model r = 0.40 + 0.48; 2nd model r = 0.35 + 0.42; 3rd model r = 0.24 + 0.10; $F_{2,11}$ = 0.19, p = 0.82). Performance of the second eye blink technology, IM System's Blinkometer, was problematic in another way. Due to problems with the Blinkometer's data storage and retrievability functions data were lost for 8 subjects. However, the remaining 6 subjects had an average bout-to-bout coherence for lapse frequency (r = 0.57 + 0.27) in the range found for the EEG algorithms and head position metrics.

Finally, although the biobehavioral technologies tested in this study were developed to detect drowsiness, they may not have detected the consequences of drowsiness (e.g., vigilance performance lapses), and therefore a relatively low coherence with PVT lapses is not evidence of their lack of validity (to detect drowsiness). This hypothesis has merit to the extent that increasing numbers of PVT lapses (or any other performance criterion) were not well correlated

with the biological state of drowsiness induced by the 42-hr sleep deprivation protocol. It is undoubtedly a safe assumption that in the current study some drowsiness episodes did not result in PVT lapses, and some PVT lapses were not due to drowsiness. However, there is no evidence to suggest that the majority of PVT lapses recorded were not due to drowsiness, since PVT lapses have been consistently demonstrated to be highly sensitive to reduced alertness associated with acute total sleep deprivation (Dinges et al., 1994), with cumulative partial sleep loss (Dinges et al., 1997): with disorders of excessive sleepiness (Kribbs & Dinges, 1994; Samuel et al., 1996; Dinges et al., in press), with circadian phase (Dinges & Kribbs, 1991; Wyatt et al., 1997), and with night work (Rosekind et al., 1994; Geer et al., 1995). Key outputs from other vigilance-based tasks documented to be sensitive to sleep and circadian pressures (antecedents of drowsiness) would also qualify as validation criteria. What will not qualify, however, is the output of a given technology's drowsiness algorithm-at least not until it has been clearly validated against some criterion relevant to drowsy driving. Consequently, technologies that purport to measure a correlate of the internal state called "drowsiness," but not its consequences as reflected in performance, must establish some other index of their validity for having relevance to the performance-impairing effects of drowsiness. Prima facie arguments for their inherent or promising validity based on anecdote, or on their biological origin (e.g.,'brain), or on their complexity or simplicity, or on authoritative claims, are not adequate. The danger of drowsy driving resides in its increased probability that the driver will fail to perform adequately to avoid a crash (Pack et al., 1995). More validation studies of this type are needed to sort out from the wide variety of biobehavioral fatigue monitors those that have the highest validity and reliability for predicting actual hypovigilance performance. This study revealed that although most technologies have potential in this regard, only one of those tested (i.e., PERCLOS) actually

performed at a high level of sensitivity and specificity, suggesting that it is a prime candidate for transition to, and validation in the driving environment.

II. EXPERIMENTAL STUDY OF EFFECTS OF ALERTING STIMULI

SPONSORED BY: NATIONAL HIGHWAY TRAFFIC SAFETY ADMINISTRATION

INTRODUCTION

The deployment of on-line driver monitoring technologies to prevent drowsy driving necessarily involves a drowsiness-detection system coupled with alerting stimuli, not only to warn the driver of hypovigilance but also to help the driver overcome the drowsiness long enough to depart the roadway and rest. Experiment I dealt with finding a valid and reliable drowsiness-detection system. Experiment II involved an analysis of the effects on PVT performance lapse frequency and PERCLOS scores of a combination of auditory and vibrotactile alerting stimuli.

Although it is known that both automobile drivers (Maycock, 1996) and commercial motor vehicle operators (Star Mountain Inc., 1997) report using fresh air, sound (e.g., radio), motor activity (e.g., chewing), and social stimulation to alert themselves when drowsiness occurs while driving, it is currently not known whether more systematically delivered alerting stimuli can reduce the occurrence of drowsiness and its consequences for hypovigilance. Therefore, in Experiment II, four subjects repeated validation Experiment I, but this second time, throughout the 42-hr period of waking, they were exposed to both auditory and vibrotactile alerting stimuli applied during the PVT performance trials.

While the original intent of Experiment II was to deliver alerting stimuli based on

feedback that drowsiness was present from one of the specific technologies that was tested in Experiment I, this was not possible at the time of Experiment II because only PERCLOS yielded the high validity and reliability needed to combine drowsiness detection with alerting stimuli, but it lacked an on-line, real-time drowsiness index readout to trigger stimuli. Therefore, Experiment II on the effects of alerting stimuli focused on determining the extent to which auditory and vibrotactile stimuli of the kind that might be deployable in a moving motor vehicle, could reduce PVT lapses across a 42-hr period of waking compared to the PVT lapse profile of the same four subjects in Experiment I (when no aIerting stimuli were present). An ancillary analysis of changes in PERCLOS-based drowsiness scores induced by alerting stimulation was also undertaken, given the high validity and reliability PERCLOS demonstrated in Experiment I.

Auditory and vibrotactile alerting stimuli were delivered during each 20-minute PVT performance bout in Experiment II, based on the average lapse profile of subjects in Experiment I (e.g., more stimuli were delivered at times when lapsing was elevated in Experiment I, such as in the middle of the night). Thus, although alerting stimuli were not contingent on actual PVT lapses (there were a number of reasons why this was neither practical nor theoretically optimal), alerting stimuli did increase in frequency and diversity as lapsing would be expected to characteristically increase across hours awake and time of day. Each vibrotactile stimulus was delivered through the hand held PVT response box. Auditory stimuli consisted of three different pre-recorded messages presented by a female voice (“stay awake, stay alert,” “‘watch for the stimulus,” and "please pay attention”).

METHODS

STUDY DESIGN

The within-subject design used in this alerting (A) Experiment II capitalized on the technology validation experiment (Experiment I, which we now refer to as the non-alerting [NA] experiment). By using four of the same subjects that participated in Experiment I, it was possible to get an estimate of the extent to which alerting stimuli delivered during the PVT performance task could be effective in reducing the magnitude and time course of PVT lapses (observed in Experiment I) during 42-hr of wakefulness. To achieve this, subjects who participated in Experiment I served as their own controls and were studied a second time in the 42-hr waking protocol, but with alerting stimuli applied during the PVT performance trials. This within-subjects design permitted comparisons of their PVT lapses profiles between the original non-alerting condition (Exp. I) and the subsequent alerting condition (Exp. II).

SUBJECTS

Four healthy subjects (ages 22-36 years) who participated in NA Experiment I were selected to run in the alerting protocol. These four were subjects 6000,6001,6011, and 60 19 (see Table 1 in Exp. I Results section). Volunteers were selected for participation based on interest, availability, and eligibility, which limited the study to 4 of the original 14 subjects involved in Exp. I. While n = 4 is a limited sample, it is important to keep in mind that each subject is an independent replication (i.e., non-alerting condition vs. alerting condition), in which the large number of repeated PVT test bouts within each condition (NA vs. A) permit statistical comparisons of the effects of alerting stimulation within each individual subject. Viewed from this perspective, Experiment II is actually comprised of four experiments.

The four subjects were healthy, nonsmokers who consumed no more than an average amount (i.e., 500mg or less) of caffeine per day. They were re-screened to confirm stable sleep/wake cycles and to be free of any sleep disorders. The 42-hr, IRB-approved alerting protocol was explained to subjects during a telephone screening session. A laboratory screening session occurred approximately l-week prior to Exp. II, at which time subjects were provided with exact details of the protocol, informed consent was obtained, and a confidential medical screen was administered to ensure that the medical state of the subject had not changed since participation in Exp. I, 4 - 7 months earlier.

PROCEDURES

Sleep History Prior to Exp. I(NA) and Exp. II (A)

At the laboratory screening session l-week prior to laboratory participation in Exp. II, subjects received a wrist actigraph and sleep diary to record sleep-wake times, which were also logged in on a voice mailbox just prior to bed and upon awakening each day, as in Exp. I. At this time, subjects were provided with an individualized specific sleep schedule that they were asked to keep for the week prior to the in-lab Exp. II. The sleep schedule provided was prepared individually for each subject based on the subject's actual sleep-wake times for the 1 week prior to non-alerting Exp. I. It was stressed to subjects that they should keep the times as close as possible to the schedule but to concentrate mainly on obtaining the same amount of sleep each night as indicated on their individual sleep schedule. This would ensure that subjects would enter the alerting protocol (Exp. II) with approximately the same amount of cumulative sleep need (Dinges et al., 1997) as they had when entering the non-alerting Exp. I protocol. This is an important prerequisite to ensuring that the sleep homeostatic pressure that produces PVT lapsing during 42-hr of waking was comparable between non-alerting Exp. I

and alerting Exp. II.

An analysis of the actigraphic, sleep diary, and electronically time-stamped telephone log records indicated that there was a trend for subjects to have acquired somewhat more sleep the night before the alerting laboratory Exp. II (M = 6.64 + 0.57 hr) than they did on the night before the non-alerting Exp. I (M = 6.12 + 0.76 hr) (t = - 2.36, df = 3, p = 0.10). This difference was also reflected in the average sleep obtained for the 4-day period prior to the alerting study (M = 7.63 + 1.09 hr) and the non-alerting study (M = 6.94 + 0.19 hr), but was not statistically reliable (t = - 1.35, df = 3, p = 0.26). To the extent that subjects obtained an average of 30 min. more sleep the night before Exp. II than on the night before Exp. I, they might be expected to be more alert (i.e., fewer PVT lapses) during Exp. II--although such a difference in sleep duration has not yet been shown to affect vulnerability to sleep loss of 42-hr. If the relatively small sleep duration difference were to affect PVT lapsing, the lapse rates in the alerting condition should be lower than those in the non-alerting condition independent of the effect of stimulation. The results of Exp. II indicate this was not found.

Experiment II Protocol

At the end of 1 week of ambulatory monitoring, subjects entered the same laboratory environment and underwent the same 42-hr sleep deprivation protocol and comparable procedures as in NA Exp. I. Thus, as in NA Exp. I, throughout the 42-hr period without sleep for A Exp. II, subjects were required to perform the same 1-hr computerized neurobehavioral assessment test battery (NAB), which contained the same 20-min. psychomotor vigilance task (PVT) (Dinges & Powell, 1985) that again served as the criterion performance task. The performance test bouts were again completed every 2 hr throughout the study (including the CTT

every other bout), as in NA Exp. I. All testing activities were performed in the same testing room as used in Exp. I, under the same lighting, sound, and staffing conditions. Once again, subjects were restricted from the consumption of any caffeine products and limited to passive activity as they were in non-alerting Exp. I. Trained staff members (biobehavioral, technical, EEG) remained with subjects at all times during the 42-hr TSD protocol of Exp. II, and as in Exp. I applied the same monitoring procedures to keep subjects awake during and in-between performance test bouts. As in NA Exp. I, throughout the 42-hr waking protocol of A Exp. II, subjects underwent monitoring of EEG and EOG, and their faces were monitored continuously throughout performance test bouts using a low light camera. Circadian phase was again. monitored by recording sublingual temperature every 2 hr, at the end of each performance bout.

In terms of hardware that came in contact with the subjects, the only procedural differences between the alerting Exp. II protocol and the earlier non-alerting Exp. I protocol were that in the alerting experiment, subjects did not wear either the MTI Research, Inc. or the IM Systems, Inc. eye-blink monitors; and in the alerting study they wore headphones during performance bouts to receive the auditory alerting stimuli (only delivered during the PVT task), and the hand-held PVT response box had an additional small attachment to deliver the vibrotactile stimulation.

Alerting Stimuli

Auditory and vibrotactile stimuli of the kind that might be deployable in a moving motor vehicle were developed and used as alerting stimuli within each PVT trial across the 42-hr period of waking in Exp. II. Technical monitors administered alerting stimuli to subjects during each 20-min. PVT bout at predetermined times, according to the schedule displayed in Table 14.

Presentation and type of alerting stimuli was selected based on the results of the non-alerting Exp. I. That is, the frequency of auditory and vibrotactile stimuli in each PVT trial of alerting Exp. II was based on an evaluation of the average profile of PVT lapsing from NA Exp. I (see Figure 2 in Exp. I). Thus, for early PVT bouts 1 and 2 (i.e., trials at 10 a.m. and noon, after 2 hr and 4 hr of waking, respectively), when lapsing was at its lowest rate in NA Exp. I, only one stimulus (vibrotactile) was delivered to subjects in Exp. II, at the 10th min. of the PVT trial (see Table 14). In contrast, during later PVT bouts 11, 12, and 13 (i.e., trials at 6 a.m., 8 a.m., and 10 a.m., after 22 - 26 hr of waking), when lapsing was at its highest rate in NA Exp. I, five alerting stimuli (auditory and vibrotactile) were delivered to subjects in Exp. II, at the 2nd, 6th, 10th, 14th, and 18th min. of the PVT trial (see Table 14). Thus, although alerting stimuli were not contingent on actual PVT lapses (there were a number of reasons why this was neither practical nor theoretically optimal), alerting stimuli did increase in frequency and diversity as lapsing would be expected to characteristically increase across hours awake and time of day.

Vibrotactile stimuli. Each vibrotactile stimulus had a duration of 5 sec. and was presented at specific times during each 20-min. PVT bout (Table 14). A small vibrotactile box was attached to the PVT box that the subject held in his hands during the PVT test. It was decided that holding the vibrotactile box in the hands would best simulate a vibration that might be given in an automobile/truck through the steering wheel.

Auditory stimuli. Auditory stimuli consisted of three different pre-recorded (audiotaped) messages presented by a female voice. The messages were: (1) "stay awake, stay alert," (2) "watch for the stimulus," and (3) "please pay attention." The subjects received these pre-recorded messages through headsets that were worn throughout all portions of the testing bout, but messages were only delivered during specified times in the PVT tests (see Table 14).

Table 14. Type and timing of auditory and vibrotactile alerting stimuli delivered during each 20-min. PVT performance test bout in Study II.

Bout	Minute	Type	Bout	Minute	Type	Bout	Minute	Type
1, 2	1		3, 5, 7	1		4, 6	1	
	2			2			2	
	3			3			3	
	4			4			4	
	5			5			5	
	6			6			6	
	7			7			7	
	8			8			8	
	9			9			9	
	10	Vibrotactile		10	"Watch for the Stimulus"		10	Vibrotactile
	11			11			11	
	12			12			12	
	13			13			13	
	14			14			14	
	15			15	Vibrotactile		15	"Watch for the Stimulus"
	16			16			16	
	17			17			17	
	18			18			18	
	19			19			19	
	20			20			20	

Bout	Minute	Type	Bout	Minute	Type	Bout	Minute	Type
8	1		9	1		10	1	
	2			2			2	
	3	"Please Pay Attention"		3	Vibrotactile		3	"Stay Awake/Stay Alert"
	4			4			4	
	5			5			5	
	6			6			6	
	7			7			7	
	8			8			8	
	9	Vibrotactile		9	"Please Pay Attention"		9	Vibrotactile
	10			10			10	
	11			11			11	
	12			12			12	
	13			13			13	
	14			14			14	
	15	"Stay Awake/Stay Alert"		15	"Stay Awake/Stay Alert"		15	"Please Pay Attention"
	16			16			16	
	17			17			17	
	18			18			18	
	19			19			19	
	20			20			20	

Bout	Minute	Type	Bout	Minute	Type	Bout	Minute	Type
11	1		12	1		13	1	
	2	"Watch for the Stimulus"		2	"Watch for the Stimulus"		2	"Stay Awake/Stay Alert"
	3			3			3	
	4			4			4	
	5			5			5	
	6	Vibrotactile		6	Vibrotactile		6	Vibrotactile
	7			7			7	
	8			8			8	
	9			9			9	
	10	"Stay Awake/Stay Alert"		10	"Stay Awake/Stay Alert"		10	"Please Pay Attention"
	11			11			11	
	12			12			12	
	13			13			13	
	14	Vibrotactile		14	Vibrotactile		14	Vibrotactile
	15			15			15	
	16			16			16	
	17			17			17	
	18	"Please Pay Attention"		18	"Please Pay Attention"		18	"Watch for the Stimulus"
	19			19			19	
	20			20			20	

Table 14 (continued)

Bout	Minute	Type	Bout	Minute	Type	Bout	Minute	Type
14	1		15, 20	1		16	1	
	2	"Please Pay Attention"		2	Vibrotactile		2	"Watch for the Stimulus"
	3			3			3	
	4			4			4	
	5			5			5	
	6	Vibrotactile		6			6	Vibrotactile
	7			7	"Please Pay Attention"		7	
	8			8			8	
	9			9			9	
	10	"Watch for the Stimulus"		10			10	"Stay Awake/Stay Alert"
	11			11			11	
	12			12	Vibrotactile		12	
	13			13			13	
	14	Vibrotactile		14			14	Vibrotactile
	15			15			15	
	16			16			16	
	17			17	"Stay Awake/Stay Alert"		17	
	18	"Stay Awake/Stay Alert"		18			18	"Please Pay Attention"
	19			19			19	
	20			20			20	

Bout	Minute	Type	Bout	Minute	Type	Bout	Minute	Type
17	1		18	1		19	1	
	2	Vibrotactile		2	"Please Pay Attention"		2	"Please Pay Attention"
	3			3			3	
	4			4			4	
	5			5			5	
	6			6			6	
	7	"Stay Awake/Stay Alert"		7	Vibrotactile		7	Vibrotactile
	8			8			8	
	9			9			9	
	10			10			10	
	11			11			11	
	12	Vibrotactile		12	"Stay Awake/Stay Alert"		12	"Stay Awake/Stay Alert"
	13			13			13	
	14			14			14	
	15			15			15	
	16			16			16	
	17	"Please Pay Attention"		17	Vibrotactile		17	Vibrotactile
	18			18			18	
	19			19			19	
	20			20			20	

RESULTS

Despite the fact that a total of 66 alerting stimuli (n = 36 auditory and n = 30 vibrotactile) were delivered once every 2 - 10 min. to each subject during the twenty PVT performance bouts across 42-hr of waking in Exp. II, no subject evidenced even a trend for a difference in bout-to-bout lapse frequency. Figures 12, 13, 14, and 15 (subjects 6000, 6001, 6011, 6019, respectively) display the PVT bout-to-bout lapse profiles for lapse frequency for non-alerting Exp. I and alerting Exp. II. Figure 16 displays the average lapse data from NA Exp. I and A Exp. II, for the

four subjects. Both profiles for individual subjects and the average data are remarkably similar between Exp. I and II despite: (1) the presence of alerting stimuli in Exp. II; (2) the modest pre-experimental sleep duration differences (see Subject section above); and (3) the 4 - 7 months that separated Exp. I and Exp. II. There is an apparently reproducible "fingerprint" quality to the overall bout-to-bout profile of PVT lapses for each subject between experiments I and II.

Table 15 displays the results of statistical comparisons (paired t-tests) between non-alerting Exp. I and alerting Exp. II for the mean number of PVT lapses across each 42-hr period of waking; for the mean PVT median reaction time (msec) across each 42-hr period of waking; and for the mean post-PVT visual-analog sleepiness rating made by subjects across each 42-hr period of waking. Only two statistically reliable differences were found and these were for data from the same subject (60 19), but did not involve PVT lapses and were internally contradictory. Thus, subject 6019 had a significantly longer PVT median RT in the alerting Exp. II ($t = 3.16$, $p = 0.0051$), but reported significantly less subjective sleepiness in Exp. II ($t = -5.30$, $p = 0.0001$). The subjective sleepiness differences between the Exp. I and Exp. II for this subject were evident even during the first few PVT trials, suggesting that this subject used the visual analog sleepiness scale differently in the alerting study relative to the non-alerting study.

PVT lapse data and PERCLOS P80 data were also compared between Exp. I and II for changes during the precise minute of the PVT in which either an auditory or vibrotactile stimulus was delivered, as well as subsequent minutes. Consistent with the lack of differences in PVT lapses across the 42-hr period of waking, there was no evidence in these comparisons that either auditory or vibrotactile stimuli had any effect on PVT lapse frequency. A third analysis evaluated changes in PVT lapses and PERCLOS P80 in each of 3-minutes post-stimulus relative the minute prior to stimulation (i.e., data from Exp. II only). This analysis suggested that for the

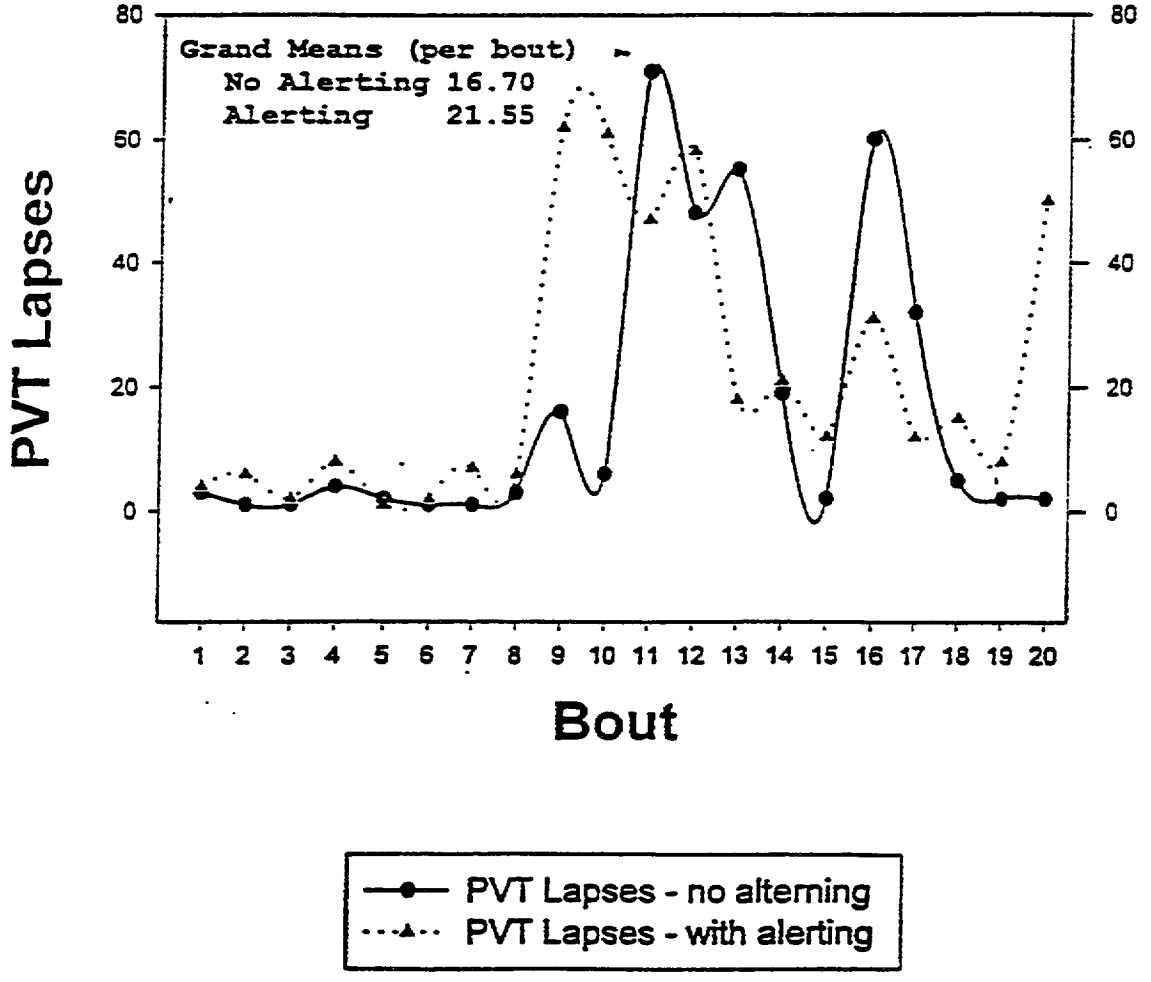

Figure 12. Bout-to-bout PVT lapse profiles for the alerting (dotted line) and non-alerting (solid line) conditions for subject 6000.

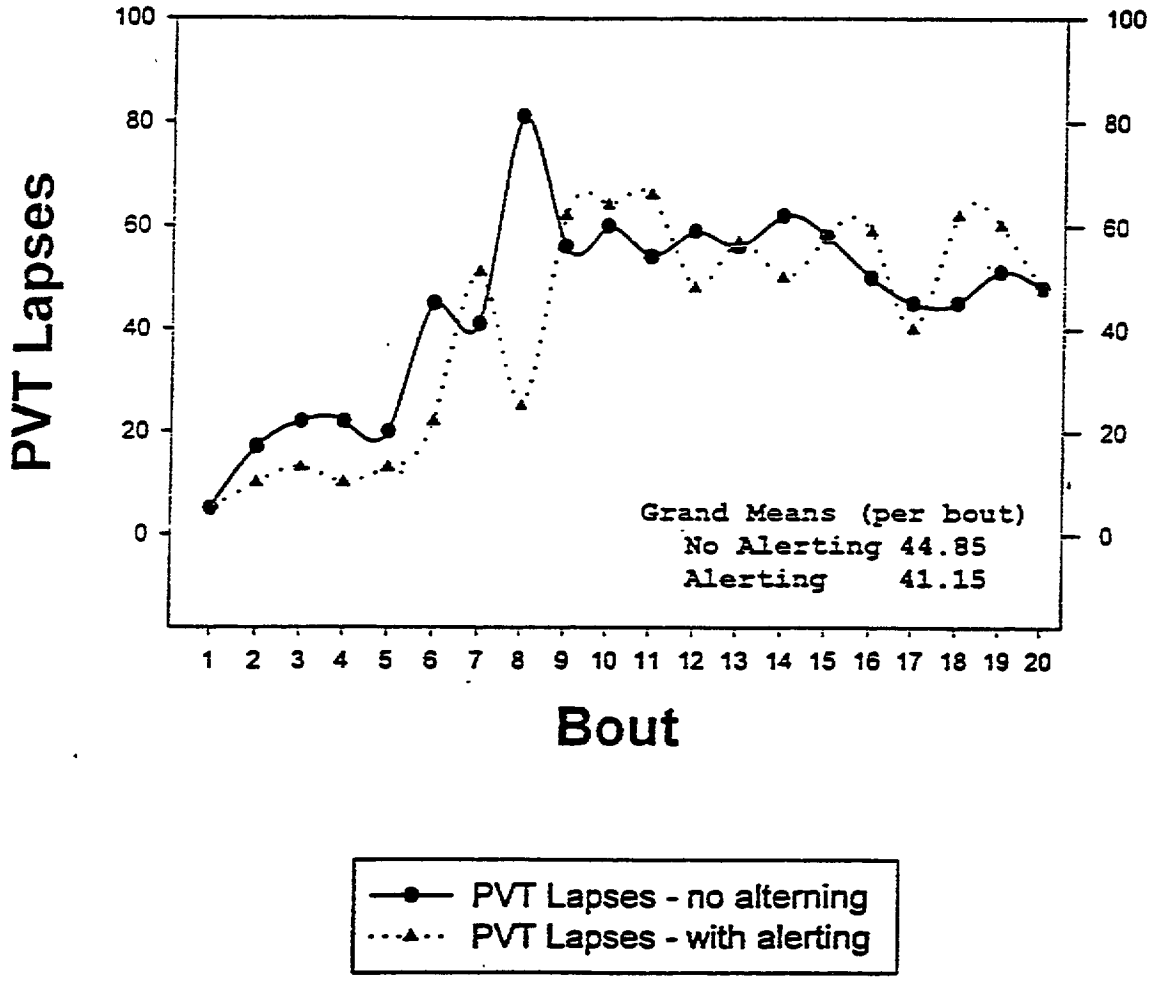

Figure 13. Bout-to-bout PVT lapse profiles for the alerting (dotted line) and non-alerting (solid line) conditions for subject 6001.

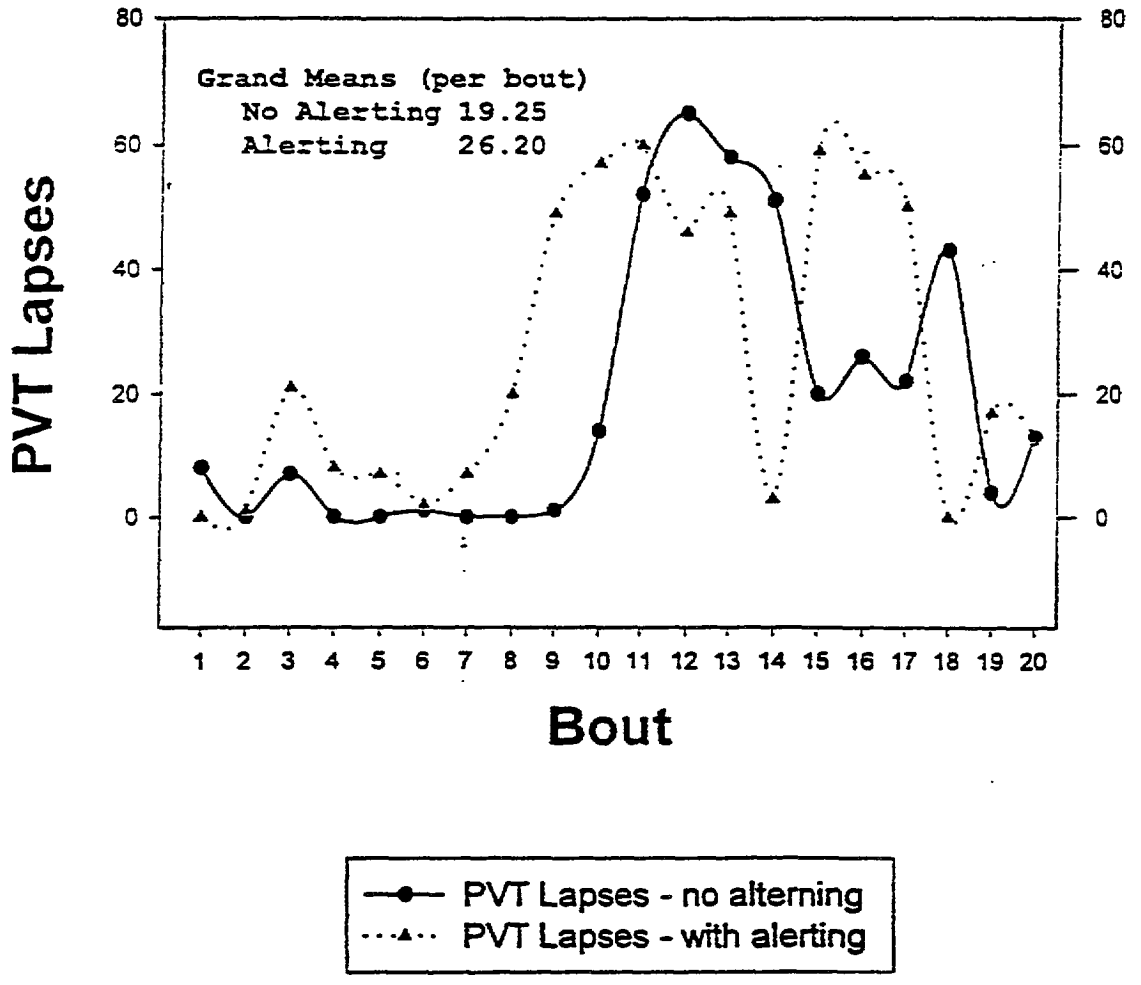

Figure 14. Bout-to-bout PVT lapse profiles for the alerting (dotted line) and non-alerting (solid line) conditions for subject 6011.

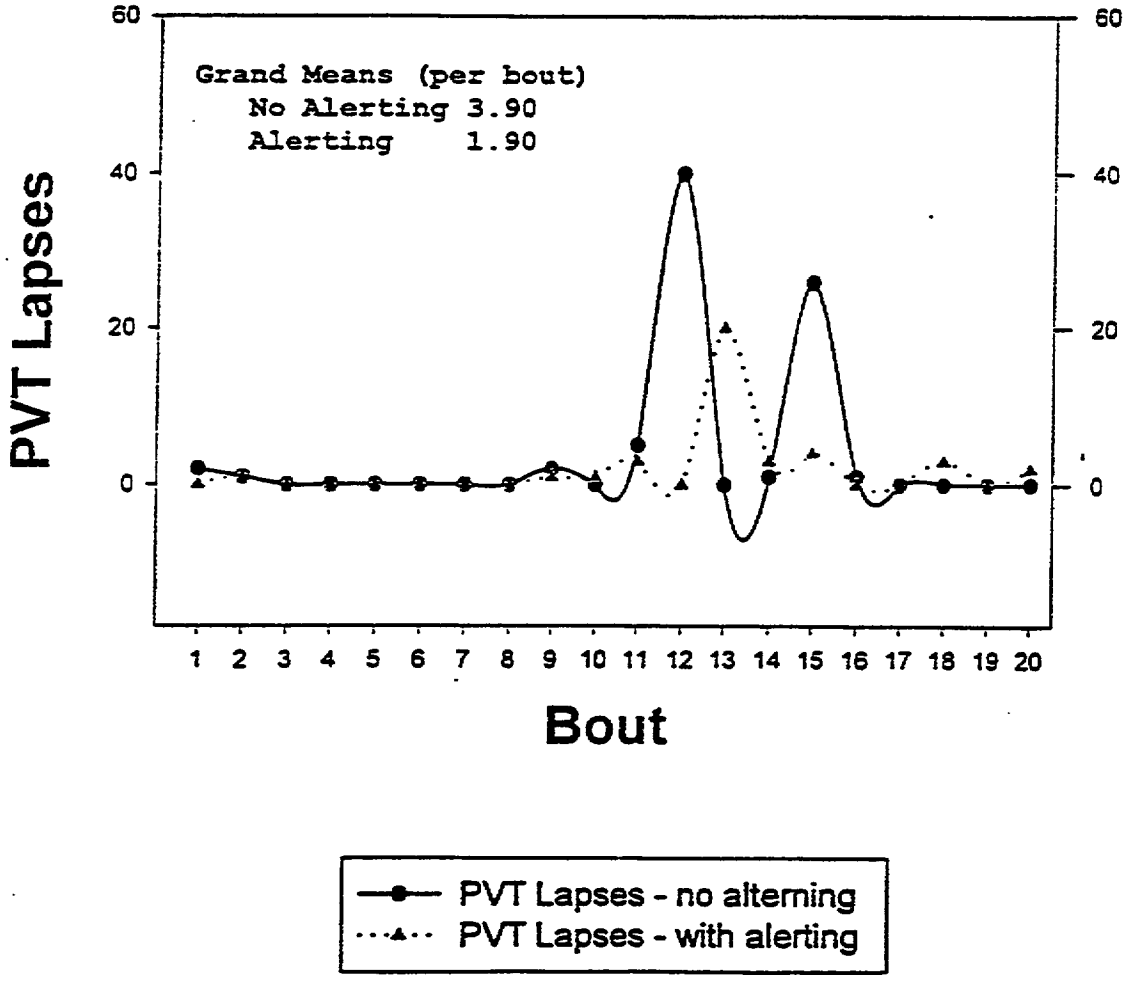

Figure 15. Bout-to-bout PVT lapse profiles for the alerting (dotted line) and non-alerting (solid line) conditions for subject 6019.

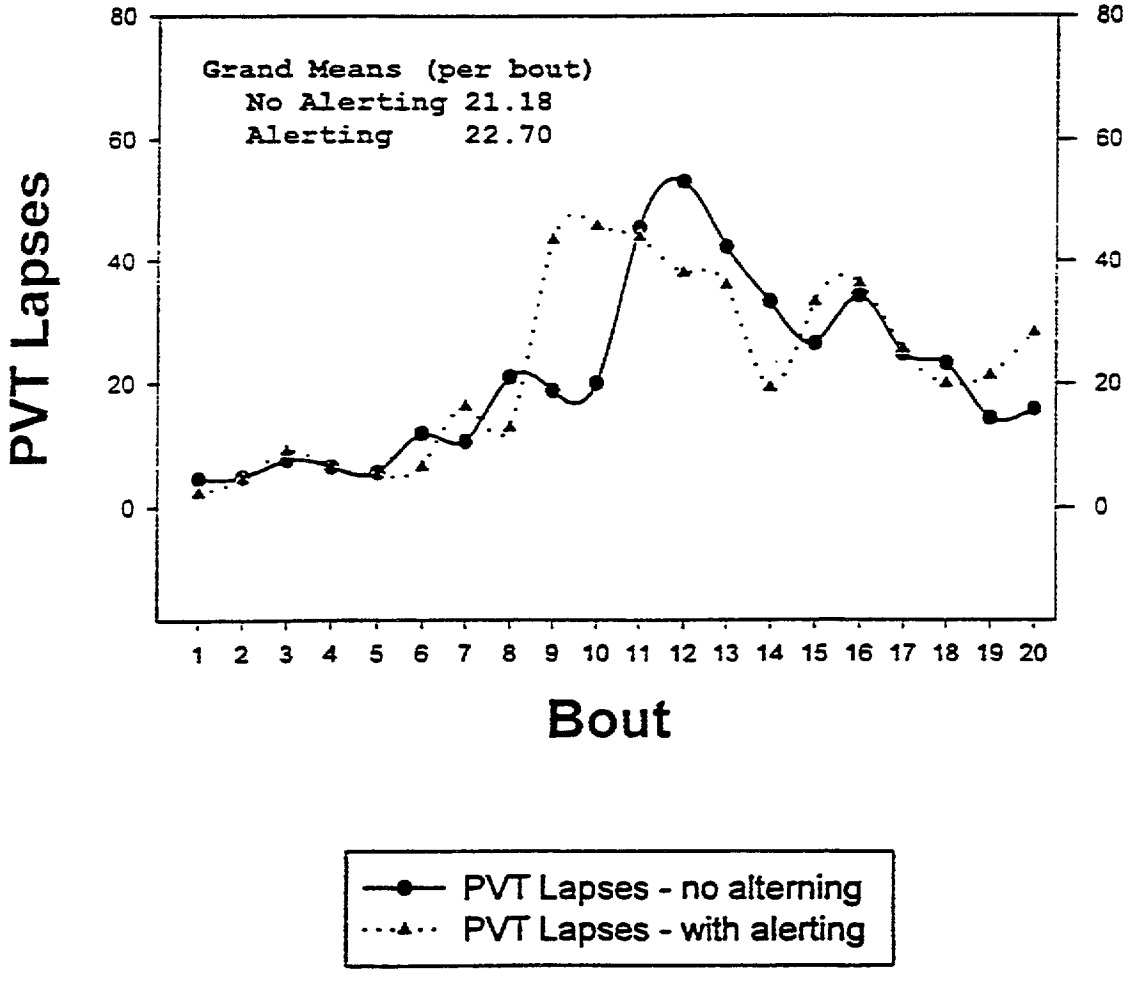

Figure 16. Bout-to-bout PVT lapse profiles for the alerting (dotted line) and non-alerting (solid line) conditions across all subjects (n=4).

minute in which the stimulus was delivered and 1 minute thereafter, there was a tendency for lapses to decrease, but this trend was not evident 2 - 3 minutes post-stimulus. This transient effect of alerting stimulation appeared to be slightly more robust for auditory stimuli than for vibrotactile stimulation-

Table 15. Comparisons (paired t-test) between non-alerting (NA) and alerting (A) conditions for 3 PVT variables from 4 subjects studied across 42-hr waking in both NA Exp. I and A Exp. II.

	P VT lapses			***PVT median***			***Post-PVT sleepiness***		
ID#	NA	A	t =	NA	A	t	NA	A	t =
6000	16.7	21.5	0.91 (ns)	1007.4	294.4	-0.99 (ns)	4.6	3.8	-1.01 (ns)
6001	44.8	41.1	-1.04 (ns)	538.4	1041.1	0.77 (ns)	7.7	7.5	-0.50 (ns)
6011	19.2	26.2	1.23 (ns)	290.9	547.6	1.61 (ns)	6.5	7.1	0.67 (ns)
6019	3.9	1.9	-0.79 (ns)	228.7	246.2	3.16 (0.0051)	3.8	1.5	-5.30 (0.0001)
mean	21.1	22.7	0.58	516.3	532.3	0.06 (ns)	5.6	5.0	-1.10 (ns)

Finally, analyses of PERCLOS results showed no effect of stimulation in any minute, replicating the lack of effects of alerting stimuli on PVT performance. However, PERCLOS again showed high validity for predicting vigilance lapses. Three of the subjects evaluated in the non-alerting validation study were also studied in this second experiment on the effects of alerting stimuli. Bout-to-bout coherence for lapse frequency between PVT lapses and PERCLOS was again very high for each subject (subject 6000, validation study coherence r = 0.917 vs. alerting study r = 0.834; subject 6001, validation study coherence r = 0.825 vs. alerting study r = 0.883; subject 6011, validation study coherence r = 0.972 vs. alerting study r = 0.917). In addition, PERCLOS coherence was calculated for subject 6019 from the alerting study

(PERCLOS coherence was not calculated for this subject in the validation study due to the subject wearing the safety glasses for MTI, see Table 5). Subject 6019 had a coherence of r = 0.956 in the alerting study, which is consistent with the high coherence values observed for 10 other subjects in the validation study, as well as with the replication of PERCLOS coherence for the three subjects run in both studies. The consistently high coherence of PERCLOS in both the technology validation study and alerting experiment, and in the studies of Wierwille and colleagues (1994), further confirms the potential utility of PERCLOS as a drowsiness detection technology. Given its consistently high coherence with states of hypovigilance during performance demands, PERCLOS should serve as the criterion measure for determining the effects of alerting stimuli on driver vigilance and safety.

EXPERIMENT II: DISCUSSION AND CONCLUSIONS

Comparisons of PVT lapses in each minute prior, during, and following individual and collective stimulation revealed that providing auditory + vibrotactile alerting stimuli did not markedly reduce lapses in drowsy subjects beyond the minute in which the alert occurred. Parallel analyses of median PVT reaction time, post-PVT subjective sleepiness ratings, and PERCLOS scores during the PVT all confirmed the basic lack of fmdings on PVT lapses. It is possible that the use of a convenience testing sequence, and therefore lack of counterbalancing of the order of completing the non-alerting (Exp. I) and alerting (Exp. I) protocols, could have obscured the effects of alerting stimulation, if the second time subjects experienced 42-hr sleep loss was more fatiguing due to loss of novelty. However, this would not explain the failure to observe effects from alerting stimulation in Exp. II when the pre-stimulus minute was used as the control/comparison. Moreover, the PVT lapse profiles between Exp. I and Exp. II look

remarkably similar (Figures 12,13,14,15,16) despite the trend for somewhat more sleep prior to Exp. II, which should have favored finding an effect from alerting stimuli. The "fingerprint" quality of the PVT lapse profiles between Exp. I and II suggests that the alerting stimuli used were far less important in the pattern of hypovigilance than characteristics of the individual subjects.

The results suggest that a study of the effects of alerting stimuli on drowsy drivers should focus on even most robust combinations of drowsiness alarms and the most potent alerting stimuli. There is some evidence that certain olfactory, thermoregulatory, and social stimuli may possess more potent alerting potential than auditory and vibrotactile stimuli. Making the drowsiness alarm contingent on drowsiness detection may also facilitate the effectiveness of alerting stimuli. These options need to be the focus of future research targeted at identifying alerting stimuli that produce the maximum duration of alertness in drowsy drivers, even if that duration is limited to only, 5, 10, or 15 minutes. Even a modest increase in the duration of alerting effect can afford critically needed time for a driver to leave the roadway and utilize a drowsiness countermeasure with an even longer time constant (e.g., waking break, exercise, nap, caffeine).

Regardless of the duration of the acute effects of alerting stimuli, there is also a need to determine how drowsiness alarms and alerting stimuli are used by drivers under time pressure. It remains to be determined whether concerns are justified that deployment of such systems in motor vehicles will encourage some drivers to continue driving in an impaired state due to the false sense of security the technology affords, or to coercion from employers, or to willful misuse of the technology (Evans, 1991; Brown, 1997). Misuse of drowsiness-driving technology is a legitimate concern along with a range of other legal/ethical issues, but a technology that can

potentially enhance safety and save lives should not be prejudged based on speculations about user behaviors (Dinges, 1995b, 1997; Dinges & Mallis, in press). In addition to determining the most effective alerting stimuli, research is needed on the manner in which people use drowsy-driving detectors.

APPENDIX

REANALYSIS USING PVT LAPSE DURATION AS THE CRITERION VARIABLE

INTRODUCTION

Included in this appendix are the results from statistical analyses that involved a second PVT criterion variable, lapse duration, for calculating coherence. The lapse duration criterion variable included cumulative (total) time in a lapse for each minute of each 20-minute PVT bout. These analyses were performed in addition to the results reported using the PVT criterion variable, lapse frequency, documented in the main body of the report, to determine whether use of the PVT lapse frequency criterion distorted the sensitivity of each drowsiness metric/algorithm tested. Data were reanalyzed using the cumulative-time-in-lapse PVT criterion (i.e., non-responsivity while the stimulus is present), to establish the extent to which very long-duration lapses (e.g., 10 sec to 30 sec duration) may have prevented the PVT lapse frequency criterion from reflecting the true state of drowsiness as detected by one or more of the technologies / algorithms. This is best illustrated by an example in which a l-minute period during the PVT task contains 2 lapses, but each is nearly 30-sec in duration, and they therefore consume all of the 1 -minute period; in contrast to a 1 -minute period during the PVT task that contains 6 brief (< 1 sec) PVT lapses. The former might be regarded as 'microsleeping" during the entire 1-min period, while the latter could be argued to be a milder dorm of drowsiness during effortful performance. Consequently, it was important to establish to what extent the findings of the validation experiment were altered by using a PVT criterion that reflected the total time lapsing more so than the frequency of lapses.

While the rationale behind the use of the lapse duration criterion variable, are sound, the reanalysis made little difference to the outcomes and conclusions in the body the report. The

results reported below for PVT lapse duration criterion confirm the results obtained using the originaI PVT lapse frequency criterion, and further reinforce the conclusion that PERCLOS outperformed all other drowsiness metrics, yielding the highest bout-to-bout and minute-to-minute coherence out of all technologies/algorithms tested.

RESULTS

BOUT-TO-BOUT COHERENCE

Bout-to-bout coherence for lapse duration refers to the correlation between the total time spent lapsing in each 20-min. PVT bout and the results of a given drowsiness detection algorithm from a given technology. Table 16 displays the mean (SD) and median bout-to-bout coherence for lapse duration for each technology. The average bout-to-bout coherence for lapse duration for eye/facial ratings of PERCLOS remained well above the average coherence for lapse duration for all other technologies. This is consistent with the reported results using the lapse frequency criterion variable (see Table 6). While all the PERCLOS metrics remained well above the metrics of EEG algorithms, head position metrics and eye blink metrics, P80 had the highest average bout-to-bout coherence using the lapse duration criterion (r = 0.91 + 0.08).

Table 16. Average bout-to-bout coherence for lapse duration (Pearson correlation coefficients).

	Eye/facial ratings			EEG algorithms		Head position metrics		Eye blink monitors	
	P70	P80	EM	CRI	SMM	ASC60	ASC90	MTI	IM
N	10	10	10	4	4	5	5	14	6
Minimum	0.64	0.71	0.78	0.45	0.14	-0.52	0.16	-0.47	0.19
Maximum	0.96	0.97	0.96	0.86	0.92	0.86	0.80	0.92	0.87
Median	0.92	0.94	0.92	0.51	0.79	0.80	0.62	0.32	0.79
Mean	0.88	0.91	0.90	0.58	0.66	0.46	0.51	0.32	0.65
Std Dev	0.09	0.08	0.06	0.19	0.35	0.59	0.30	0.35	0.27

MINUTE-TO-MINUTE COHERENCE

Minute-to-minute coherence' for lapse duration refers to the correlation between the total time spent lapsing in each minute of a 20-min. PVT bout across the 42-hr of waking, and the results of a given drowsiness detection algorithm from a given technology. Table 17 displays the mean (SD) and median minute-to-minute coherence for lapse duration for each technology. As with bout-to-bout coherence, the average minute-to-minute coherence for lapse duration for the eye/facial ratings of PERCLOS remained well above the average coherence for lapse duration for all other technologies tested (Table 9). As with bout-to-bout coherence, P80 had the highest average minute-to-minute coherence (r = 0.76 + 0.10) for the lapse duration criterion.

Table 17. Average minute-to-minute coherence for lapse duration (Pearson correlation coefficients).

	eye/facial ratings			*EEG-algorithms*		*Head position metrics*		*Eye blink monitors*	
	P70	**P80**	**EM**	**CRI**	**SMM**	**ASC60**	**ASC90**	**MTI**	**IM**
N	10	10	10	4	4	5	5	14	6
Minimum	0.47	0.56	0.59	0.36	0.22	-0.20	0.002	-0.28	0.25
Maximum	0.88	0.89	0.88	0.65	0.76	0.60	0.43	0.72	0.62
Median	0.75	0.78	0.78	0.41	0.68	0.35	0.24	0.28	0.46
Mean	0.71	0.76	0.76	0.46	0.59	0.27	0.21	0.26	0.43
Std Dev	0.12	0.10	0.09	0.13	0.24	0.31	0.16	0.27	0.14

Minute-to-minute coherence for lapse frequency was consistently lower than bout-to-bout coherence for lapse frequency (see Table 10). This result was also observed in the lapse duration analysis. These comparisons are displayed in Table 18. Thus, regardless of PVT criterion variable, nearly all technologies had a statistically significantly better prediction of PVT performance when sampling involved a longer (i.e., 20 min.) rather than a briefer (i.e., 1 min.) time period.

Table 18. Comparison of bout-to-bout and minute-to-minute coherence measures for lapse duration (Pearson correlation coefficients).

	Eye/facial ratings			*EEG algorithms*		*Head position metrics*		*Eye blink monitors*	
	P70	P80	EM	CRE	SMM	ASC60	ASC90	VET	IM
Number Ss studied	10	10	10	4	4	5	5	14	6
By Bout 1 – 20	0.88	0.91	0.90	0.58	0.66	0.46	0.51	0.32	0.65
By Minute 1 - 20	0.71	0.76	0.76	0.46	0.59	0.27	0.21	0.26	0.43
Mean Difference	-0.17	-0.15	-0.14	-0.12	-0.07	-0.19	-0.30	-0.06	-0.22
t =	-9.6	-8.6	-8.0	-3.2	-1.3	-1.3	-3.1	-2.0	-3.4
p =	0.0001	0.0001	0.0001	0.04	ns	ns	0.03	0.06	0.01

COHERENCE VARIABILITY

Day 1 vs. Day 2 of Waking

In order to determine whether each bout-to-bout coherence for lapse duration was comparable across the broad range of sleepiness/alertness induced by the within-subjects experimental design, coherence was also recalculated separately for the first 22-hr of wakefulness (i.e., performance bouts 1 to 10; from 10:00 a.m. on day 1 to 4:00 a.m. on day 2), and compared to coherence for the final 20-hr of waking (i.e., performance bouts 11 to 20; from 6:00 a.m. on day 2 to just after midnight on the start of day 3), when subjects were much sleepier and lapses increased in duration. Table 19 displays the results of these analyses for both bout-to-bout and minute-to-minute coherence for lapse duration. The findings are very similar to those for lapse frequency criterion (see Table 1 l), with the exception of the PERCLOS variables (P70, P80, EM). For PVT lapse frequency criterion, there was a trend for coherence to be higher during the first 22hr awake than during the final 20hr awake (Table 11). Using the lapse duration criterion, these differences evaporated and the average PERCLOS coherence was comparably high (i.e., r = 0.83 – 0.90) for both early and later periods of wakefulness (Table 19). Shifting to

a lapse duration criterion resulted in a mean increase in PERCLOS coherence for bouts I l-20 (e.g., P80 mean lapse frequency coherence for bouts 1 l-20, r = 0.65; compared to P80 mean lapse duration coherence for bouts 1 l-20, r = 0.87), but the increase was not statistically significant (t = - 1.66, p = 0.13).

It is also noteworthy that minute-to-minute coherence values increased for PERCLOS metrics when the lapse duration criterion was used relative to the lapse frequency criterion (compare Table 1 I to Table 19 minute-to-minute coherence values). However, the increase was greater for the final 20 hr of waking than for the first 22 hr of waking, which resulted in statistically significant differences in PERCLOS coherence for lapse duration between the first 22 hr awake and the final 20 hr awake (see Table 19 bottom half). PERCLOS minute-to-minute mean coherence for the first 22 hr awake (e.g., P80 r = 0.55), the final 20 hr awake (P80 r = 0.69), and the total 42 hr awake (P80 r = 0.76) was the highest achieved by any technology/algorithm when lapse duration was used as the criterion (Tables 17 and 19).

Table 19. Coherence measures for lapse duration for bouts #1 to 10 vs. bouts #11 to 20 (Pearson correlation coefficients).

	Eye/facial ratings			*EEG algorithms*		*Head position metrics*		*Eye blink monitors*	
	P70	P80	EM	CRT	SMM	ASC60	ASC90	MET	TM
Number Ss Studied	10	10	10	4	4	5	5	14	6
By Bout									
1-10 (1st 22hr awake)	0.88	0.90	0.85	0.28	0.41	0.39	0.43	0.18	0.13
11 - 20 (2nd 20 hr awake)	0.83	0.87	0.87	0.46	0.75	0.47	0.39	0.13	0.37
Paired t-test t=	-1.2	-0.8	0.4	0.6	0.8	0.6	-0.5	-0.4	0.8
p=	ns	ns	ns	ns	ns	ns	ns	ns	ns
By Minute									
1 - 10 (1st 22 hr awake)	0.51	0.55	0.55	0.22	0.43	0.20	0.20	0.17	0.22
11 - 20 (2nd 20 hr awake)	0.65	0.69	0.71	0.43	0.63	0.20	0.14	0.23	0.29
Paired t-test t=	2.2	2.4	2.8	1.9	1.1	0.0	-1.2	0.7	0.4
P -	0.05	0.04	0.02		ns	ns	ns	ns	ns

COMPARISON OF COHERENCE FOR LAPSE FREQUENCY VS. LAPSE DURATION

The percent declines in coherence going from bout-to-bout measures to minute-to-minute measures was greatly diminished when using the PVT criterion variable, lapse duration (instead of lapse frequency). While bout-to-bout coherence improved for some technologies, the improvement in minute-to-minute coherence was substantial in many cases. The most significant improvement, out of all the technologies tested, in the minute-to-minute range using lapse duration was displayed by PERCLOS. The results of the statistical comparisons of bout-to-bout coherence for lapse frequency criterion versus lapse duration criterion are displayed in Table 20. Minute-to-minute coherence comparisons are shown in Table 2 1.

Table 20. Bout-to-bout coherence measures for lapse frequency vs. lapse duration (Pearson correlation coefficients).

	Eye/facial ratings			*EEG algorithms*		*Head position metrics*		*Eye blink monitors*	
	P70	**P80**	**EM**	**CRI**	**SMM**	**ASC60**	**ASC90**	**MTI**	**IM**
Number Ss studied	10	10	10	4	4	5	5	14	6
Lapse frequency	0.86	0.87	0.87	0.53	0.62	0.46	0.52	0.33	0.57
Lapse duration	0.88	0.91	0.90	0.58	0.66	0.46	0.51	0.32	0.65
Mean Difference	-0.02	-0.04	-0.03	-0.05	-0.04	0.005	0.01	0.01	-0.08
t =	-0.88	-1.55	-0.97	-0.88	-0.30	-0.16	0.73	0.43	-1.75
p =	ns	ns	ns	ns	ns	ns	ns	ns	ns

Table 21. Minute-to-minute coherence measures for lapse frequency vs. lapse duration (Pearson correlation coefficients).

	Eye/facial ratings			*EEG algorithms*		*Head position metrics*		*Eye blink monitors*	
	P70	**P80**	**EM**	**CRI**	**SMM**	**ASC60**	**ASC90**	**MTI**	**IM**
Number *Ss studied*	10	10	10	4	4	5	5	14	6
Lapse frequency	0.61	0.63	0.63	0.29	0.46	0.30	0.26	0.22	0.29
Lapse Duration	0.71	0.76	0.76	0.46	0.59	0.27	0.21	0.26	0.43
Mean Difference	*-0.1*	*-0.13*	-0.13	-0.17	-0.13	0.03	0.05	-0.04	-0.14
t =	-1.82	-2.29	-2.03	-1.66	-1.01	0.89	-1.81 ,	-1.47	-2.76
P =	0.101	0.047	0.072	ns	ns	ns	ns	ns	0.039

Changing the PVT criterion from lapse frequency to lapse duration did not result in any statistically significant increases in bout-to-bout coherence for any of the technologies / algorithms (Table 20). In contrast, the change in PVT criterion significantly increased minute-to-minute coherence for PERCLOS variable P80, with a trend for improvement in P70 and EM metrics, as well as in minute-to-minute coherence for IM System's Blinkometer. The most striking outcome was the fact that with lapse duration criterion, P80 minute-to-minute coherence averaged r = 0.76, which places this variable closer to the optimal coherence achieved for bout-to-bout using lapse frequency (i.e., mean r = 0.87) than to minute-to-minute coherence for lapse frequency (i.e., mean r = 0.63). Thus, P80 minute-to-minute coherence for PVT lapse duration was comparable to its lapse fi-equency coherence for 4-minute and 5-minute windows (see Figure 10).

DISCUSSION AND CONCLUSIONS

Despite a change in the lapse criterion used to calculate coherence, PERCLOS continued to yield the highest correlation with psychomotor vigilance performance. Thus, the new PVT criterion variable, lapse duration, did not change the rank order among the various drowsiness metrics of each technology tested. PERCLOS continued to display the highest coherence when using lapse duratiqn for both minute-by-minute and bout-by-bout coherence measures, and it's minute-to-minute coherence under the lapse duration criterion reached a level close to that found for bout-to-bout coherence. The Alertness Monitor continued to show low coherence with the lapse duration criterion variable, and the Blinkometer, head position and EEG algorithms still achieved only mid-range coherence values, although Dr. Makeig's drowsiness metric had the second highest coherence value next to all PERCLOS metrics. The analyses using the new PVT criterion variable (lapse duration) show that PERCLOS was a more complete predictor of

cumulative time in lapse than other technologies'/ algorithms. These results fully confirm the results of the original study, and further reinforce the conclusion that PERCLOS shows the greatest promise as an on-line drowsiness detection modality at this time.

REFERENCES

ACGIH Standards, 1995 edition

Akerstedt, T, Gillberg, M: Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience 52: 29-37,1990*.

Babkoff, H, Caspy, T, Mikulincer, M, Sing, HC.: Monotonic and rhythmic influences: a challenge for sleep deprivation research. *Psychol.* Bull. 109(3):41 l-428, 1991.

Bloomfield, P: *Fourier analysis of time series,* New York: John Wiley & Sons, p. 214, 2976.

Brillinger, DR: *Time series: data analysis and theory, expanded edition, San* Francisco: Holden-Day, p.257, 198 1.

Brown, ID: Methodological issues in driver fatigue research. *In* Hartley, L. (Ed) *Fatigue & Driving: Driver impairment, Driver Fatigue and Driving Simulation,* Taylor & Francis, London, 1995.

Brown, ID: Prospects for technological countermeasures against driver fatigue. *Accident Analysis and Prevention 29: 525-53 1, 1997.*

Brookhuis, K.: Driver impairment monitoring by physiological measures in Hartley, L. (Ed) *Fatigue & Driving: Driver Impairment, Driver Fatigue and Driving Simulation,* Taylor & Francis; London, 1995.

Dinges, DF, Powell, JW: Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments and Computers 17:652-655,* 1985.

Dinges, DF, Ome, MT, Whitehouse, WG, Ome, EC: Temporal placement of a nap for alertness: Contributions of circadian phase and prior wakefulness. Sleep 10:3 13-329, 1987.

Dinges, DF, Powell, *JW:* Sleepiness is more *than* lapsing. *Sleep Research* 17:84,1988.

Dinges, DF: The nature of sleepiness: Causes, contexts and consequences. In Stunkard, A., Baum, A. (Eds) *Perspectives in Behavioral Medicine: Eating, Sleeping and Sex,* Lawrence Erlbaum, Hillsdale, 1989.

Dinges, DF, Graeber, RC: Crew fatigue monitoring. *Flight Safety Digest 8: 65-75,1989.*

Dinges, DF, Kribbs, NB: Performing while sleepy: Effects of experimentally-induced sleepiness. In Monk, T.H. (Ed.) *Sleep, Sleepiness and Performance,* John Wiley and Sons, Ltd., Chichester, United Kingdom, 199 1.

Dinges, DF, Kribbs,, NB, Steinberg, KN, Powell, JW: Do we lose the willingness to perform during sleep deprivation? *Sleep Research 21:3 18,* 1992.

Dinges, DF, Kribbs, NB, Bates, BL, Carlin, MM: A very brief probed-recall memory task: Sensitivity to sleep loss. *Sleep Research 22: 330, 1993.*

Dinges, DF, Douglas, SD, Zaugg, L, Campbell, DE, McMann, JM, Whitehouse, WG, Ome, EC, Kapoor, S.C., Icaza, E., Ome, M.T.: Leukocytosis and natural killer cell function parallel neurobehavioral fatigue induced by 64 h of sleep deprivation. The *Journal of Clinical Investigation 93:* 1930-l 939, 1994.

Dinges, DF: Technology / Scheduling Approaches in Managing Fatigue in Transportation: Promoting Safety and Productivity. In *Proceedings@om the multi-modal symposium, co-*sponsored by the National Transportation Safety Board and NASA Ames Research Center, 53-58, 1995a.

Dinges, DF: An overview of sleepiness and accidents. *Journal of Sleep Research 4:* (Supplement 2); 4-14, 1995b.

Dinges, D.F: Validation of psychophysiological monitors. In *Proceedings of the Technological Conference on Enhancing Commercial Motor Vehicle Driver Vigilance, American* Trucking Associations Foundation, FHWA, NHTSA, McLean, VA, 35-41,1996.

Dinges, DF: The promise and challenges of technologies for monitoring operator vigilance. In *Proceedings of the International Conference on Managing Fatigue in Transportation,* American Trucking Associations Foundation, Tampa, FL, 77-86, 1997.

Dinges, DF, Pack, F, Williams, K, Gillen, KA, Powell, JW, Ott, GE, Aptowicz, C, Pack, AI: Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4-5 hours per night. *Sleep 20(4).-267-277, 1997.*

Dinges, DF, Mallis, MM: Managing fatigue by drowsiness detection: Can technological promises be realized? *In* Hartley, L. (Ed) *Coping with the 24 Hour Society: Fatigue Management Alternatives to Prescriptive Hours of Service* from the Proceedings of the 3rd International Conference on Fatigue in Transportation, Taylor & Francis, Bristol, Pennsylvania, in press.

Dijk, D-J, Duffy, JF, Czeisler, CA: Circadian and sleep/wake dependent aspects of subjective alertness and cognitive performance. J: *Sleep Res.1:* 112-l 17,1992.

Evans, L: *Traffic Safety and the Driver.* Van Nostrand Reinhold: New York, 1991.

Fukuda, J, Adachi, K, Nishida, M, Akutsu, E: Development of driver's drowsiness detection technology. *Toyota Technical Review 45: 34-40,* 1995.

Geer, RT, Jobes, DR, Gilfor, J, Traber, KB, Dinges, D: Reduced psychomotor vigilance in anesthesia residents after *24-hr* on-call. *Anesthesiology* 83: (Supplement 3A), A1008, 1995.

Hamelin, P: Lorry drivers' time habits in work and their involvement in traffic accidents. *Ergonomics 30.* 1323-33,1987.

Harris, W: Fatigue, circadian rhythm and truck accidents. In Mackie, *RR* (Ed) *Vigilance: Theory, Operational Performance and PhysiologicaI Correlates.* Plenum Press, New York, 1977.

Hoddes, E, Zarcone, V, Smythe, H, Phillips, R., Dement, WC: Quantification of sleepiness: a new approach. *Psychophysiol., 10: 431-436, 1973.*

Home, JA, Reyner LA: Sleep related vehicle accidents. *British Medical Journal 310: 565-567,1995.*

Jung T-P, Makeig S, Stensmo M, Sejnowski TJ: Estimating alertness from the EEG power spectrum. *IEEE Transactions on Biomedical Engineering 44: 60-69,1997.*

Kaneda, M, Iizuka, H, Ueno, H, Hiramatsu, M, Taguchi, M, Tsukino, M: Development of a Drowsiness Warning System. *Proceedings of the 14th International Technical Conference on Enhanced Safety of Vehicles,* Munich, Germany, l-7,1994.

Knipling, RR, Wang, JS: Revised estimates of the US drowsy driver crash problem size based on general estimates system case reviews. *In 39th Annual Proceedings,* Association for the Advancement of Automotive Medicine, Chicago, 1995.

Knipling, R: The promise of technology for fatigue management: The Federal Highway Administration perspective. In *Proceedings of the Technological Conference on Enhancing Commercial Motor Vehicle Driver Vigilance, American* Trucking Associations Foundation, FHWA, NHTSA, McLean VA, 8-l 3, 1996.

Kribbs, N B Dinges, DF: Vigilance decrement and sleepiness. In Harsh, JR & Ogilvie, RD (Eds.), *Sleep* Onset *Mechanisms. American* Psychological Association. Washington, DC, pp. 113-125,1994.

Makeig, S, Inlow, M: Lapses in alertness: coherence of fluctuations in performance and EEG spectrum. *Electroencephalography and Clinical Neurophysiology 86 :23-35,I993.*

Makeig, S, Elliott, FS, Postal, M: First Demonstration of an alertness monitoring management system; Report No. 93-36, Naval Health Research Center, San Diego, CA, l-20,1993.

Makeig, S, Jung, T-P: Tonic, phasic, and transient EEG correlates of auditory awareness in drowsiness. *Cognitive Brain Research&* (1):15-25, Jul 1996.

Makeig, S, Jung ,T-P.: Changes in alertness are a principal component of variance in the EEG spectrum. *NeuroReport 7: 2* 13-2 16 ,1995.

Maycock, G: Sleepiness and driving: The experience of U.K. drivers. *Accident Analysis and Prevention, 29,: 453-462,1997.*

Mitler, MM, Carskadon, MA, Czeisler, CA, Dement, WC, Dinges, DF, Graeber, RC: Catastrophes, sleep and public policy: Consensus report of a committee for the Association of Professional Sleep Societies. *Sleep 11:* 100-109, 1988.

McNair, DM, Lorr, M, Druppleman, LF: *EITS Manual For the Profile of Mood States. San* Diego, Educational and Industrial Test Services, 1971.

McDonald, N: Fatigue, Safety and the Truck Driver. Taylor & Francis: London, 1984.

0' Hanlon, JF, Kelley, GR: *A psychophysiological evaluation of devices for preventing lane drift and run-off-road accidents,* Federal Highway Administration, Report No. 1736- F, 1974.

O'Hanlon, JF: What is the extent of the driving fatigue problem? In *Driving Fatigue in Road Traffic Accidents,* Brussels: Commission of the European Communities Report No. EUR6065EN, l9-25,1978.

Pack, AI, Pack, AM, Rodgman, E, Cucchiara, A, Dinges, DF, Schwab, CW: Characteristics of crashes attributed to the driver having fallen asleep. *Accident Analysis and Prevention* 27: 769-775, 1995.

Rau, P: The National Highway Traffic Safety Administration's Drowsy Driver Technology Program. In *Proceedings of the Technological Conference on Enhancing CommerciaI Motor Vehicle Driver Vigilance,* American Trucking Associations Foundation, FHWA, NHTSA, McLean, VA, 14-16, 1996.

Richardson, JH: The development of a driver alertness monitoring system. In Hartley, L (Ed) *Fatigue & 'Driving: Driver Impairment, Driver Fatigue and Driving Simulation.* Taylor & Francis; London, 219-229, 1995.

Rosekind, MR, Graeber, RC, Dinges, DF, Connell, LJ, Rountree, MS, Spinweber, CL, Gillen, *KA,.: Crew factors in flight operations: LX Effects of cockpit rest on crew performance and alertness in long-haul operations.* NASA Technical Memorandum Rep. No. 103884, 252pp, 1994.

Rowland, L, Thome, D, Balkin, T, Sing, H, Wesensten, N, Redmond, D, Johnson, D, Anderson, A, Cephus, R, Hall, S, Thomas, M, Powell, J, Dinges, D, Belenky, G: The effects of four different sleep-wake cycles on psychomotor vigilance. *Sleep Research 26: 627, 1997.*

Samuel, S., Pack, FM, Pack, AI, Maislin, G, Winokur, A, Dinges, DF: Sleep-wake behaviors of elderly living in residential communities: results from a *case-control* study of excessive *daytime* sleepiness. *Sleep Research 25:* 13 1, 1996.

Shafer, JH: The decline of fatigue related accidents on *the NYS* thruway. *In Proceedings of the Highway Safety Forum on Fatigue, Sleep Disorders and Traffic Safety,* Albany, NY, 1993.

Star Mountain, Inc.: *Survey of Truck Drivers to Determine Knowledge and Beliefs Regarding Driver Fatigue.* Revised Analysis report prepared for the Trucking Research Institute, January 21,1997.

Stem, JA: Blink rate: a possible measure of fatigue. *Human Factors, 36 (2):285-297,* 1994.

Thayer, RE.: Activation-deactivation adjective checklist: current overview and structural analysis. *Psychological Reports,* 58:607-614, 1986.

U.S. National Transportation Safety Board Safety Study: *Fatigue, Alcohol, Other Drugs, and Medical Factors in Fatal-to-the-Driver Heavy Truck Crashes,* Vol. 1, Washington, D.C., NTSB/SS-90-01,1990.

Wang, JS, Knipling, *RR: Single-Vehicle roadway departure crashes: Problem size assessment and statistical description.* US Department of Transportation, National Highway Traffic Safety Administration, 1994.

Wierwille, *WW,* Ellsworth, L.A.: Evaluation of driver drowsiness by trained raters, *Accident analysis and Prevention, 26* (5):571-581, 1994.

Wierwille, WW, Ellsworth, LA, Wreggit, *SS,* Fairbanks, *RJ, Kim,* CL: *Research on vehicle-based driver status/performance monitoring: development, validation, and refinement of algorithms for detection of driver drowsiness.* National Highway Traffic Safety Administration Final Report: DOT HS 808 247, 1994.

Wyatt, J.K., Dijk, D-J., Ronda, J.M., Jewett, M.E., Powell, J.W., Dinges, D-F., Czeisler, CA.: Interaction of circadian-and sleep/wake homeostatic-processes modulate psychomotor vigilance test *(PVT)* performance. *Sleep Research 26: 759, 1997.*

Wylie, CD, Shultz, T, Miller, JC, Mitler, *MM,* Mackie, *RR: Commercial Motor Vehicle Driver Fatigue and Alertness Study: Project Report,U.S. Department of Transportation Report* No. FHWA-MC-97-002,1996.

DOT HS 808 762
April 1998

People Saving People
http://www.nhtsa.dot.gov